2017

Multi-omic biomarker discovery and network analyses to elucidate the molecular mechanisms of lung cancer premalignancy

BOSTON UNIVERSITY

GRADUATE SCHOOL OF ARTS AND SCIENCES

AND

COLLEGE OF ENGINEERING

Dissertation

MULTI-OMIC BIOMARKER DISCOVERY AND NETWORK ANALYSES TO ELUCIDATE THE MOLECULAR MECHANISMS OF LUNG CANCER PREMALIGNANCY

by

ANNA MARIA TASSINARI

B.S., University of Massachusetts, 2009
M.S., Boston University, 2014

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2017

Approved by

First Reader _____

Jennifer Beane, Ph.D.

Assistant Professor of Medicine

Second Reader _____

Avrum Spira, M.D., M.Sc.

Professor of Medicine

DEDICATION

To my husband Nate, for your unconditional love, boundless compassion, and disarming humor I could not live without.

To Mom, for your unwavering encouragement, inspiration, and excellence I will forever admire and aspire to achieve.

ACKNOWLEDGMENTS

This work would not have been possible without:

My advisors and mentors – Jen, Avi and Marc: Thank you for your enthusiasm and passion for science you all shared with me. I will forever be grateful for the trust you've placed in me and the independence you've granted me. Thank you for providing me with room to grow and an opportunity to learn from my mistakes.

The Computational Biomedicine Group: Thank you for teaching me something new every day, fostering critical thinking and exchanging ideas, and inspiring discussions.

The Bioinformatics Program: Thank you for taking the chance and giving me the unique opportunity to learn from so many brilliant scientists in the field. And thank you for always having my back.

My family - Mom and Dad Krentowicz: Thank you for your love which has not lessened despite the 4,000 miles between us. I miss you.

My extended family - Jo and Dan Tassinari: Thank you for welcoming me into your family with open arms and creating a home away from home for me.

My friends - Thank you for restoring my sense of belonging and all the memories we've created together. I will always cherish them.

My rock - Nate Tassinari: there are no words to describe my gratitude for your care, support and encouragement throughout this journey. You are the rock. You are the sea.

MULTI-OMIC BIOMARKER DISCOVERY AND NETWORK ANALYSES TO ELUCIDATE THE MOLECULAR MECHANISMS OF LUNG CANCER PREMALIGNANCY

ANNA MARIA TASSINARI

Graduate School of Arts and Sciences, and College of Engineering, 2017

Major Professors: Jennifer Beane, Assistant Professor of Medicine and Avrum Spira, Professor of Medicine

ABSTRACT

Lung cancer (LC) is the leading cause of cancer death in the US, claiming over 160,000 lives annually. Although CT screening has been shown to be efficacious in reducing mortality, the limited access to screening programs among high-risk individuals and the high number of false positives contribute to low survival rates and increased healthcare costs. As a result, there is an urgent need for preventative therapeutics and novel interception biomarkers that would enhance current methods for detection of early-stage LC.

This thesis addresses this challenge by examining the hypothesis that transcriptomic changes preceding the onset of LC can be identified by studying bronchial premalignant lesions (PMLs) and the normal-appearing airway epithelial cells altered in their presence (i.e., the PML-associated airway field of injury). PMLs are the presumed precursors of lung squamous cell carcinoma (SCC) whose presence indicates an increased risk of developing SCC and other subtypes of LC. Here, I leverage high-

throughput mRNA and miRNA sequencing data from bronchial brushings and lesion biopsies to develop biomarkers of PML presence and progression, and to understand regulatory mechanisms driving early carcinogenesis.

First, I utilized mRNA sequencing data from normal-appearing airway brushings to build a biomarker predictive of PML presence. After verifying the power of the 200-gene biomarker to detect the presence of PMLs, I evaluated its capacity to predict PML progression and detect presence of LC (Aim 1). Next, I identified likely regulatory mechanisms associated with PML severity and progression, by evaluating miRNA expression and gene coexpression modules containing their targets in bronchial lesion biopsies (Aim2). Lastly, I investigated the preservation of the PML-associated miRNAs and gene modules in the airway field of injury, highlighting an emergent link between the airway field and the PMLs (Aim 3).

Overall, this thesis suggests a multi-faceted utility of PML-associated genomic signatures as markers for stratification of high-risk smokers in chemoprevention trials, markers for early detection of lung cancer, and novel chemopreventive targets, and yields valuable insights into early lung carcinogenesis by characterizing mRNA and miRNA expression alterations that contribute to premalignant disease progression towards LC.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

cfDNA	cell-free DNA
CIS	Carcinoma *in situ*
CT	Computer Tomography
ctDNA	circulating tumor DNA
DNA	Deoxyribonucleic Acid
FDR	False Discovery Rate
FET	Fisher's Exact Test
GSEA	Gene Set Enrichment Analysis
miRNA	microRNA
NLST	National Lung Screening Trial
NSCLC	Non-Small Cell Lung Cancer
PC	Principal Component
PCA	Principal Component Analysis
PCGA	Pre-Cancer Genome Atlas
PML	Premalignant Lesion
RNA	Ribonucleic Acid
SCC	Squamous Cell Carcinoma
TCGA	The Cancer Genome Atlas
WGCNA	Weighted Gene Co-expression Network Analysis

CHAPTER ONE: Introduction

1.1. Tobacco-Induced Lung Cancer As The Leading Cause Of Cancer Death

Every 200 seconds someone in the U.S. loses their battle with lung cancer [51]. While the great majority of these deaths could be prevented by eliminating their main cause - the active exposure to tobacco smoke, lung cancer could be cured even in long-time heavy smokers if it was detected early, i.e. before it has spread to lymph nodes or metastasized. In fact, only about 18% of patients live with the diagnosis more than 5 years - a number shockingly low in comparison to breast cancer's (90%), prostate cancer's (99%), and colorectal cancer's (65%) 5-year survival rates [51]. Yet, there exists a significant disparity among these cancers in research funding. In 2014, the National Cancer Institute awarded an estimated $254mln (https://fundedresearch.cancer.gov) to lung cancer research, paying more than twice as much to breast cancer which accounts for only a quarter as many annual deaths [51]. Counterintuitively, lung cancer's societal burden, which can be quantified as years of life lost (YLL) or disability-adjusted life-years (DALY) significantly surpasses that of other leading cancers, contributing to increased health-care and economic costs [27]. Motivated by the low survivability and high societal costs, many researchers have focused their efforts on developing much needed early lung cancer screening methods, that would allow detection in subjects presenting without evident symptoms.

In 2002-2004 the National Lung Screening Trial (NLST) recruited over 50,000 participants with smoking history to compare the efficacies of low-dose computed-tomography (CT) and chest radiography (X-ray) in reducing mortality from lung cancer by detecting tumors at an early stage [109]. Participants were randomized into two groups, screened annually for 3 years with the group-assigned method and followed. The NLST Research Team found that low-dose CT performed better at detecting clinically significant abnormalities and ultimately led to a reduction in deaths 15-20% larger as compared to X-ray. However, the rate of overdiagnosis by CT (percentage of suspected lung tumors ending up being slow-growing and non-life threatening or completely benign) of non-small cell lung cancer (NSCLC) was high, at 22.5% [32,91]. These results suggest that benefits from early lung cancer detection with CT should be weighed against the potentially incurred extra healthcare cost as well as health risks including unnecessary surgery or harmful chemotherapy in overdiagnosed cases, and that refocusing efforts on the precancerous stages of disease where intervention benefits outweigh the risks may improve lung cancer mortality rates just as well.

1.2. Bronchial Premalignant Lesions (PMLs) And Their Role In Lung Carcinogenesis

Bronchial premalignant lesions (PMLs) are histological abnormalities in the central airways, characterized by variably thickened basement membrane separating the epithelium from the underlying connective tissue [1]. They can be observed and sampled

via autofluorescence bronchoscopy, and reproducibly graded by a pathologist [86]. The natural history of PMLs, which follows a step-wise progression model whereby normal cells proceed through pathological stages from basal-cell hyperplasia and squamous metaplasia, to mild, moderate and severe dysplasia, to carcinoma in-situ, has been well documented [63,748,103].

Collectively referred to as dysplasia, PMLs are the presumed precursors of lung squamous cell carcinoma (SCC) [55]. Several premalignant lesion prevalence studies showed that high-grade dysplasia has a higher prevalence in patients with invasive carcinoma and that the lesion number and severity correlate with increased risk for developing lung cancer [8,10,48,90,127]. What is more, subjects with lesions that progress or persist over time, have been shown to have increased risk of progressing to invasive SCC [83], and more often than not, the invasive SCC was observed to develop at a location that was different from the location of the monitored lesion [136]. In addition, there exists limited evidence that mortality from lung cancer could improve by 90% if the premalignant lesions were detected and treated early [31]. However, because of the fact that PMLs are dynamic and do not strictly follow the linear progression model (their histology might worsen and improve multiple times within a patient [22]) there is a lack of a well-established causal link between the lesions and the disease, which may explain why the current standard of care excludes monitoring PMLs in the context of prevention or early detection of lung cancer. Thus, pursuing PML detection and monitoring in the

context of early stage diagnostics may help refine the link between premalignant disease and lung tumorigenesis.

1.3. Role of Biomarkers in Disease Management

Biomarker discovery is a promising area of research focusing on deciphering predictive signatures responsible for disease onset, development, or prognosis, and applying these signatures to improve disease diagnosis, management, and treatment. The National Institutes of Health Biomarkers Definitions Working Group defines a biomarker as "a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention." [124] Depending on their intended clinical purpose, biomarkers can aid in disease-risk assessment (e.g. serum-based biomarker for Alzheimer's disease [76] or blood gene expression-based biomarker for sarcoidosis [150]), detection of preexisting conditions (e.g. lipid-based biomarker for Acute kidney injury [29,108]), determination of disease progression rate (e.g. imaging-based biomarkers for Parkinson's disease [81] and Alzheimer's disease [82]), and patient stratification for improved targeted therapy (e.g. EGFR mutations in non-small cell lung cancer [129]). In addition, many biomarker-based diagnostics use non-invasive procedures making them more accessible, easier to administer, and posing lesser patient burden.

Of great interest to this work are lung cancer biomarkers. A wide range of sample types, platforms and molecules have been used over the past several decades to develop

biological markers that would aid in the identification of subjects at high risk for developing lung cancer or those with a potentially curable early-stage disease (as reviewed in Brothers *et al.* [23]). Among epigenetic biomarkers, many focus on processes involved in DNA methylation, which can affect downstream phenotypes without the modification of the DNA itself. For example, Belinsky *et al.* have used genome-wide methylation profiling of sputum and lung biopsy samples to identify P16-INK4A (CDKN21) as hypermethylated in lung tumors and sputum samples of smokers[17]. This result was recapitulated one year later in another study that used paired serum and lung tissues [35]. In addition, high methylation of *SHOX2* was determined to be indicative of the presence of malignant disease in subjects with suspect lung cancer using bronchial fluid aspirated during bronchoscopy [106] and plasma [65]. More recently, a risk classification model was used to identify a panel of three genes whose methylation status was capable of identifying subjects at varied lung cancer risk levels (i.e., *RASSF1A* was found to be hypermethylated in the high-risk group with at least 60% chance of developing lung cancer, while *3OST2* and *PRDM14* were hypermethylated in the low-risk groups) [52]. Among gene expression-based biomarkers, some utilize easily accessible bronchial brushings of normal-appearing epithelium and the notion of field cancerization, which characterizes the transcriptomic changes in the normal airway as reflective of smoking status and lung cancer. For example, Spira *et al.* have developed an 80-gene biomarker that distinguished between samples from subjects with and without lung cancer with 83% accuracy [120]. A follow-up study has demonstrated the biomarker's independence from

clinical covariates and its increased value as a component of a combined clinicogenomic model [14]. The biomarker has since been refined [141] and validated [114]. Another study identified a panel of 14 antioxidant and DNA repair-associated genes whose extreme expression (either low or high) was prevalent in brushes from subjects with and without lung cancer [19]. In addition, several blood-based gene-expression diagnostic biomarkers have been attempted. Although white blood cells have been shown to share alternations characteristic of lung tumors [113,146], a limited number of transcriptomic studies demonstrates high performance in plasma and serum, mostly due to the instability of circulating mRNA [23]. However, several studies have now turned to liquid biopsy as a means to identify blood-based biomarkers using circulating tumor DNA (ctDNA) and cell-free DNA (cfDNA), which are believed to be released into the bloodstream via processes such as apoptosis and necrosis, and whose levels can be easily monitored in patients over time. Although still in nascent stages of development and validation, these proof-of-concept biomarkers show great promise in monitoring tumor burden [33,37,128], identification of mechanisms of resistance in EGFR-mutated NSCLC [89,117], prediction of response to chemotherapy with carboplatin in KRAS-mutated NSCLC [87], or prediction of response to tyrosine kinase inhibitor in EGFR-mutated NSCLC [77]. Finally, miRNA-based biomarkers have played a significant role in lung cancer diagnosis and assessment of premalignant disease. In a 2009 study, Mascaux *et al.* have identified a panel of miRNAs dysregulated in premalignant lesion biopsies of varied severity [79]. In another study, a small set of miRNAs were shown to discriminate between subjects with stage I SCC and

controls [143]. Interestingly, several miRNA-based biomarkers have also leveraged liquid biopsy in plasma samples. Boeri and Sozzi *et al.* have identified [21] and validated in large Italian cohort, a plasma-based miRNA panel, which when coupled with low-dose CT (LDCT), reduced the LDCT rate of overdiagnosis five-fold [118]. Furthermore, the 24-miRNA biomarker was also evaluated from a perspective of disease prognosis. At baseline, biomarker scores from 84 subjects with LC detected by LDCT were binned according to the estimated level of risk, and checked again at follow-up five years later. Survival rates were found to be strongly correlated with assigned risk groups [110]. Also utilizing plasma samples, Shen *et al*. found miR-21, miR-210 and miR-486-5p to be capable of distinguishing plasma samples from subjects with malignant vs. benign solitary pulmonary nodules (SPNs) detected previously by CT (or controls) with 75% sensitivity and 85% specificity [112]. Similarly, a panel of 13 miRNAs discovered using sputum samples, was shown to differentiate between malignant and benign SPNs and offer potential reduction in CT-related overdiagnosis rates [142].

1.4. Concepts and Methodologies

1.4.1. RNA and miRNA

Phenotypic diversity is believed to be vastly influenced by variation in gene expression patterns. An intermediary between DNA and protein, mRNA molecules (messenger RNAs) carry genetic information from the cell's nucleus to the cytoplasm, and the control of their transcription from DNA and translation into protein plays a

crucial role in the regulation of gene expression and thus cell function. Molecular events impacting the behavior of oncogenes and tumor-suppressor genes, such as mutations, gene amplifications, and chromosomal rearrangements are well described in carcinogenesis (as reviewed in Brothers *et al.*[96]). Recently, post-transcriptional regulation mediated by microRNAs (miRNAs) has been observed to play an important role in lung carcinogenesis [93]. miRNAs are short (~22 nt) non-protein-coding single-stranded RNAs (ssRNAs) which regulate gene transcription by binding via imprecise sequence-specific base-pairing to their mRNA target's 3' end. One miRNA can regulate tens to hundreds of genes at once, and it is said that 30% of protein-coding genes are affected by miRNAs [75,93]. In cancer, dysregulation of oncogenic miRNAs can promote excessive cell proliferation and impairment of apoptosis by targeting tumor suppressors for degradation, expression reduction, or both [57]. Recently, Perdomo *et al* demonstrated the importance of miR-4423 in primate-specific repression of lung carcinogenesis and regulation of airway epithelial cell differentiation [94]. miRNAs promise clinical utility as disease biomarkers and potential therapeutic targets due to their increased stability and tissue-specific expression compared to mRNAs.

1.4.2. High-Throughput Sequencing and Microarrays

The detection and quantification of gene expression is possible thanks to DNA microarrays. A microarray is a microscope slide with thousands of tiny holes containing probes of genes or other known DNA sequences organized in a grid. mRNA samples

collected from studied individuals are first converted into cDNA (complementary DNA), labeled with distinct color dyes and allowed to hybridize (bind) to the chip. The expression of each gene in a sample is treated as proportional to the observed intensity of color in a location on the chip with a given gene. Although newer technologies have started to replace microarrays in many research labs, historically these arrays have facilitated the creation of an extensive collection of differential expression studies that should not be overlooked.

More recently, next generation sequencing (NGS) has taken transcriptomics by the storm by offering a wider dynamic range of detection, the ability to identify novel transcripts, reduced background noise, and increased cost-effectiveness due to multiplexing capabilities. RNA-Seq protocol typically involves isolating total RNA from samples under investigation. For mRNA-specific sequencing, the naturally polyadenylated (poly-A) mRNA molecules are first purified using oligo-dT magnetic beads and fragmented. Then, the RNA fragments are reverse transcribed into the first strand complementary DNA (cDNA), after which the second cDNA strand is also synthesized. Fragment ends are repaired (overhangs are converted into blunt ends), and a single A nucleotide is added to the 3' end. Indexing adapters containing a complimentary T nucleotide are then ligated to the cDNA fragments. Using PCR, DNA libraries carrying adapters are amplified and then summarized to reflect overall transcript abundance (Illumina®). Because miRNAs lack the poly-A tail, additional steps have to be taken to select for these small molecules prior to reverse transcriptase. Specifically, total RNA can

be size-fractionated by gel electrophoresis, which involves cutting and purifying a gel fragment containing only sequences of desired length (~22 nucleotides) [66].

Most analyses in this thesis rely on RNA and miRNA-Seq transcriptomics experiments, but microarrays play an important role in validating of some of the findings.

1.4.3. Genomic Biomarker Development

A typical biomarker development process involves multiple stages designed to ensure proper formulation of the biomarker, its adequate validation (both from a computational as well as a clinical utility point of view), and finally translation into the clinic [41]. The general idea behind the various steps is to provide a robust and useful tool that will fulfill an unmet need, be easy to administer, and pose great benefit and minimal risk to the patient.

Briefly, at the discovery stage markers are first identified using methodologies aimed at answering a predefined question. This may include genes or miRNAs whose expression correlates with treatment response, disease prognosis, or subject risk-stratification. Ideally, the biomarker can then be internally validated on a random subset of samples left out of the biomarker discovery process to prevent bias. Alternatively, cross-validation approaches can also be employed in cases where samples size is small. In addition, at this stage, biomarker's independence from clinical and demographic covariates (e.g. sex, age, smoking status, prior cancer history) is also established. Then, external validation in additional independent datasets takes place, as well as experimental

validation of biomarker candidates. Clinical utility is then assessed by evaluating performance metrics such as sensitivity (percentage of subjects with a condition who are identified as such), specificity (percentage of subjects without a condition who are identified as such), positive predictive value (probability of a positive test correctly identifying subjects with a condition) and negative predictive value (probability of a negative test correctly identifying subjects without a condition).

1.4.4. Network Analyses in Transcriptomics

Unlike monogenic diseases (e.g. sickle-cell anemia or Huntington's), lung cancer is a complex disease, in which atypical phenotype is manifested through an abnormality not in a single gene but the entire complex molecular machinery. To understand how these anomalies cause disease phenotypes, it is essential to study the entire system, as the organization within biological networks is not random [11].

Typically, biological networks can be visually represented by graphs – mathematically derived net-like structures containing nodes connected to each other with edges. While nodes often correspond to molecular components of the cell, edges can represent a wide array of biological interactions between them. For example, metabolic networks represent the biochemical and molecular processes that take place in a cell in order to maintain life, with many subnetworks corresponding to metabolic pathways [20]. Cell signaling networks showcase how individual signaling pathways affect each other, elucidating the manner in which a biological system may respond to a signal. In protein-

protein interaction networks [56]. In epistasis interaction networks, genes are connected if there exists an interaction between them when one is knocked out or down-regulated [107]. Edges in disease-gene interaction networks represent mutational events that cause or contribute to the disease and typically connect a phenotype to a genotype underscoring the complexity of disease. In drug interaction networks, therapeutics are linked to their targets, highlighting the many-to-many relationships [88]. Finally, gene regulatory networks display cellular mechanisms governing cell function by regulation of mRNA transcription and translation into protein [60].

Of great interest to this work are gene coexpression networks, which use expression-profile correlation as a measure of gene similarity and are becoming popular tools in biomedical research [116]. Weighted Gene Coexpression Network Analysis (WGCNA) [147] described in detail in Chapter 3 has been successfully applied in late onset Alzheimer's disease (LOAD) to identify an immune and microglia module strongly associated with pathophysiology of LOAD, and *TYROBP* as a master regulator implicated in neuronal damage [18]. Another study used WGCNA to develop a small-cell lung cancer specific classifier with capacity to stratify patients for therapy. Spleen tyrosine kinase (*SYK*) was identified as a potential oncogene which could be targeted in *SYK*-positive patients [135]. In general, network-based approaches offer advantages in biological and clinical settings, as they provide system-wide context for single genes implicated in a disease and elucidate the influence of interconnectedness of cellular components on phenotype. What is more, they may aid in disease classification and drug

target identification (reviewed in Barabasi *et al.* [11]). Finally, we can use them to understand the molecular modifications that take place in a diseased or otherwise perturbed system, by exploring the topological alterations we observe in a network.

1.5. Dissertation Aims

This thesis examines the hypothesis that transcriptomic changes preceding the onset of lung cancer can be identified by studying bronchial premalignant lesions (PMLs) and the normal-appearing airway epithelial cells altered in their presence (i.e., the PML-associated airway field of injury). In the following aims, I leverage high-throughput mRNA and miRNA sequencing data from bronchial brushings and lesion biopsies to develop biomarkers of PML presence and progression, and to understand regulatory mechanisms driving early carcinogenesis.

Aim1: Develop and validate a gene expression-based biomarker for lung cancer premalignancy in the airway field of injury

Since PMLs are the presumed precursors of squamous cell carcinoma and tend to occur more frequently in patients with invasive carcinoma, they constitute risk factors for developing lung cancer. Currently detectable only via autofluorescence bronchoscopy, PMLs are not monitored as part of standard of care partially due to limited access to and relative invasiveness of the technology, as well as lack of reliable surrogate endpoints (i.e. intermediate markers such as PML regression that (a) may strongly correlate with and (b) be more easily measured in lieu of true endpoints such as a decrease in lung

cancer-related mortality). In addition, PML histology within a patient can vary greatly and more peripheral lesions may be missed by bronchoscopy. Thus, to address these challenges, in this aim I utilized mRNA sequencing data from normal-appearing airway brushings obtained from the main stem bronchus using widely-available white-light bronchoscopy, to build a biomarker predictive of PML presence, with a capacity to identify PMLs likely to progress. Given the field's potential to reflect the overall health of the airway, the biomarker could help identify high-risk smokers in need of a more aggressive follow-up.

Aim2: Identify miRNA/mRNA regulatory interactions associated with severity and progression of lung cancer premalignant lesions

Although the natural history of PMLs is thought to follow a step-wise process in which histologically normal cells gradually acquire the characteristics of cytological atypia, the PMLs are dynamic and can worsen and improve multiple times within a patient. In addition, little is known about the effects of mRNA/miRNA interactions on lesion progression. In this aim, I sought to find out if gene and miRNA expression extracted from lesion biopsies harbored information about PMLs' potential to change over time. I identified likely regulatory mechanisms associated with PML severity and progression, by evaluating miRNA expression and gene coexpression modules containing their targets in bronchial lesion biopsies.

Aim3: Identify shared genomic and regulatory alterations associated with lung carcinogenesis in the field, the lesion and the tumor.

Gene expression extracted from bronchial brushings proved to have great utility in detecting the presence of PMLs in Aim1. Additionally, in Aim 2 gene expression extracted from lesion biopsies was demonstrated to be reflective of PML severity and progression. In this aim, I examined the preservation of the PML-associated miRNAs and gene modules in the airway field of injury, highlighting an emergent link between the airway field and the PML and thus a potential for leveraging bronchial brushes to monitor PMLs over time. In addition, I evaluated the preservation of PML-associated regulatory mechanisms in tumors.

CHAPTER TWO: Developing and Validating a Gene Expression-Based Biomarker for Lung Cancer Premalignancy in the Airway Field of Injury

2.1. Background

Bronchial premalignant lesions (PMLs) are histological abnormalities in the airway characterized by the presence of dysplastic tissue [8]. Although typically presumed to be precursors for lung squamous cell carcinoma (SCC), PML presence anywhere in the airway can be a risk marker for the development of any subtype of lung cancer [54]. Several premalignant lesion prevalence studies showed that high-grade dysplasia was often present in patients with invasive carcinoma [127] and that the lesion number and severity correlate with increased risk for developing lung cancer [reviewed in Banerjee *et al.*[10]]. In addition, mortality from lung cancer is estimated to improve by 90% if the premalignant lesions were detected and treated early [10]. In light of this evidence, pursuing PML detection in the context of chemoprevention looks promising, and could help identify high-risk patients in need of such an intervention.

Although CT screening has been successful at detecting early stage lung cancers, it is not sensitive enough to detect central lesions. As an alternative, airway monitoring can be performed with the use of flexible bronchoscopy. In order to visualize the lesions, the procedure requires the use of an autofluorescent light that makes the affected bronchial mucosa appear brown in comparison to the surrounding unaffected tissue which appears green. While fairly effective, with 89% sensitivity (as opposed to 67% of white

light bronchoscopy) [36], autofluorescence bronchoscopy is not commonly available and many major health centers across the country lack the access to this equipment. A much more conventional and accessible technology is white-light bronchoscopy, which, while widely used as a follow-up diagnostic in patients with suspect lung cancer nodules detected by CT, lacks the sensitivity to detect PMLs and central cancers such as SCC . In addition, due to the lack of successful chemopreventive agents and effective risk assessment methods, the need for intermediate end-point biomarkers that would account for more than just demographic and clinical characteristics is clear [63].

Cigarette smoke is a known carcinogen [50] which induces smoking-related airway damage. Upon exposure, the genomic response of cytologically normal epithelial cells that line the respiratory tract becomes altered, reflecting the existence of an airway "field of injury". In the past couple of decades, several studies have demonstrated the validity of this hypothesis [111,119]. In 2007, Beane *et al.* have used bronchial brushings to identify subsets of gene expression changes that were slowly reversible or irreversible upon smoking cessation [15]. In the context of chemoprevention, whereby "pharmaceutical interventions slow or reverse the progression of pre-malignancy to invasive cancer" [121], in a study by Gustafson *et al.* the PI3K pathway, which plays a crucial role in cell growth and survival, was shown to be upregulated in the normal airway of subjects with dysplasia and respond to chemoprevention treatment with myo-inositol [44]. Similarly, in a placebo-controlled trial conducted by Keith *et al.* [62], treatment with iloprost (a prostacyclin found to be downregulated in lung cancer) caused an overall decrease in

average dysplasia grade across all observed lesions among former smokers. In the context of chronic obstructive pulmonary disease (COPD), Steiling *et al.* have identified transcription factor *ATF4* as a transcriptional regulator of the airway gene expression response [122], while Androutsopoulos *et al.* found that genes implicated in the metabolism of xenobiotics, including the cytochrome P450 enzyme *CYP1A1*, were upregulated in the normal airway of smokers with COPD [6]. Similarly, miRNA expression modifications have been demonstrated in the normal airway exposed to tobacco smoke [105]and linked to smoking-related medical conditions: A set of antioxidant genes was found to be upregulated in smokers with chronic bronchitis [46]. Multiple miRNAs have been identified as slowly-reversing their expression upon smoking cessation and linked to lung cancer due to their association with cell differentiation and inflammatory disease pathways [138]. Finally, miR-218 was identified as a tumor suppressor regulating inhibition of cancer cell proliferation and *EGFR*-mediated migration in NSCLC [151]. Given the strong genomic signal resulting from cigarette-smoke exposure, the airway may be a marker for not only smoking status, but also the underlying disease state that is caused by the smoking-related gene and miRNA modifications.

Spira *et al* have previously developed a gene-expression biomarker for lung cancer [120]utilizing the airway field of injury hypothesis. The biomarker was built using airway brushings from 77 patients undergoing bronchoscopy, and achieved accuracy, sensitivity and specificity of 83%, 80%, and 84%, respectively, on an independent test set. When coupled with cytopathology as compared to bronchoscopy alone, the

biomarker showed a two-fold increase in diagnostic sensitivity. The 80-gene signature predicted cancer in patients with negative cytopathology with 95% certainty; similarly, both negative tests predicted the patient to be cancer-free with 70% certainty. In addition, a follow-up study has demonstrated the biomarker's independence from clinical covariates and its increased value as a component of a combined clinicogenomic model [14]. More recently, the gene-expression based biomarker was refined and validated in a large clinical trial [114,141]. The 17-gene biomarker provides an intermediate step between a non-diagnostic bronchoscopy and invasive nodule biopsy, in patients with intermediate pulmonary nodules detected by CT. These promising results suggest that the large-scale adoption of such lung cancer diagnostic biomarkers in the clinic could reduce patient care costs and shorten the wait time before a final diagnosis is made improving outcome and patient's quality of life.

In this chapter, we are extending this paradigm by developing a gene-expression based biomarker using cytologically normal airway epithelial cells that reflect the presence of premalignant lesions in high-risk smokers. This valuable tool developed to detect and assess bronchial premalignant lesions could aid in stratification of high-risk patients into chemoprevention trials, identification of the most effective chemopreventive agent, and detection of lesions warranting aggressive follow-up.

2.1. Methods

2.1.1. Sample Collection

Bronchial brushings of normal-appearing airway collected cross-sectionally from high-risk subjects with and without PMLs profiled by RNA-Seq

Autofluorescence bronchoscopy was performed to obtain bronchial airway brushings from subjects enrolled in the British Columbia Lung Health Study (BC-LHS) at the British Columbia Cancer Agency (BCCA) (Vancouver, BC) between June 2000 and March 2011[130]. In addition, during the procedure PMLs were sampled (if present) and evaluated by a team of pathologists. Histological lesion grade was assigned to each sampled PML and the worst histology observed was recorded and assigned to the corresponding, normal-appearing brushing. The study participants with normal or hyperplasia histology enrolled in the BC-LHS were current or recent former smokers between 50 and 75 years old with no prior history of lung cancer, who smoked for at least 20 years and whose estimated 3-year lung cancer risk was at least 2%. Baseline bronchial brushes were collected from subjects with evidence of PMLs enrolled in multiple chemoprevention studies, were current or recent former smokers between 40 and 79 years old with no prior history of lung cancer and at least 30 pack years (i.e. having smoked 1 pack a day for 30 years).

Bronchial brushes of normal-appearing epithelium from 84 BCCA subjects (1 brush from each subject) with and without PMLs were selected to undergo mRNA-Seq

while ensuring balanced clinical covariates, such as age, pack years, race, sex, and COPD status.

The data is available from NCBI's Gene Expression Omnibus (GEO) using the accession ID GSE79315.

Bronchial brushings of normal-appearing airway collected longitudinally from high-risk subjects with history of PMLs profiled by RNA-Seq

Additional bronchial airway brushings were obtained from subjects participating in the High-Risk Lung Cancer-Screening Program at Roswell Park Cancer Institute (RPCI) (Buffalo, NY) between December 2009 and March 2013. These subjects were at high risk for developing lung cancer by either having a prior history of lung or aerodigestive cancers or by being a current or recent former smoker at least 50 years of age, with at least 20 smoked pack years. Fifty-one bronchial brushes of normal-appearing epithelium from 23 RPCI subjects with and without PMLs were profiled by mRNA-Seq (18 subjects had 2 procedures and 5 subjects had 3 procedures). Samples were classified as stable/progressive if the worst histological grade at the second time point for a given patient remained the same or worsened, and regressive if the worst histological grade at the second time point improved. The RPCI samples were utilized in biomarker validation to evaluate its power to identify subjects with progressing lesions by calculating differences in the biomarker score between sequential procedures. The data is available as part of GSE79315.

Bronchial brushings of normal-appearing airway collected cross-sectionally from subjects with and without COPD and PMLs profiled by microarrays

A total of 238 bronchial airway brishings from current and former smokers with and without COPD and PMLs were profiled on Affymetrix Human Gene 1.0 ST Array as described in Steiling *et al* [122] and used as an external source of validation samples to further evaluate the biomarker's ability to detect PMLs. CEL files with mRNA expression were downloaded from GEO (GSE37147) and processed by Dr. Jennifer Beane using Robust Multi-array Average (RMA) [53] and the Ensembl Gene CDF v16.0.0 file (http://brainarray.mbni.med.umich.edu/Brainarray/Database/CustomCDF/16.0.0/ensg.asp) to standardize gene annotation. Fourty four samples were filtered out based on sex mismatches and quality. A subset of 36 samples profiled on microarrays as part of the study originated from BCCA subjects and was used as an overlapping validation set. The remaining 158 bronchial airway brushings obtained and profiled in the same manner from additional subjects at high risk of developing lung cancer, were used as an independent validation set.

Bronchial brushings of normal-appearing airway collected cross-sectionally from subjects with and without lung cancer and profiled by microarrays.

Current and former smokers with suspect lung cancer underwent flexible bronchoscopy as part of two additional microarray studies. A total of 164 samples described by Spira *et al.* [120] were profiled on Affymetrix HG-U133A Array and deposited

in GEO as GSE4115. A total of 299 samples described by Silvestri *et al* [114] were profiled on Affymetrix Human Gene 1.0 ST Array and deposited in GEO as GSE66499. These extra bronchial brushing datasets were downloaded from GEO in CEL format and processed using Robust Multi-array Average (RMA) [53] and the Ensembl Gene CDF v16.0.0 file (http://brainarray.mbni.med.umich.edu/Brainarray/Database/CustomCDF/16.0.0/ensg.asp) to standardize gene annotation. Both studies were used to validate the biomarker's ability to distinguish brushings from subjects with and without lung cancer.

Tumor and adjacent normal biopsies collected cross-sectionally from subjects with lung squamous cell carcinoma and profiled by RNA-Seq.

Tumor (n=502) and matched adjacent normal (n=51) samples collected from subjects with squamous cell lung cancer were profiled by mRNA-Seq by The Cancer Genome Atlas (TCGA) Research Network Team [137]. RSEM-normalized log2-transformed counts along with the corresponding clinical data were downloaded from the UCSC Xenabrowser, which provides access to TCGA's Genomic Data Commons (GDC) Data Portal, and used to evaluate the biomarker's ability to distinguish normal from tumor samples originating from subjects diagnosed with lung cancer.

The Institutional Review Boards (IRBs) of all participating institutions approved the studies and all subjects provided written informed consent.

2.1.2. RNA Library Preparation and RNA Sequencing

Total RNA was extracted from bronchial brushings using miRNeasy Mini Kit (Qiagen). Sequencing libraries were prepared from total RNA samples using Illumina® TruSeq® RNA Kit v2 and multiplexed in groups of four using Illumina® TruSeq® Paired-End Cluster Kit. Each sample was sequenced on the Illumina® HiSeq® 2500 to generate 100-nucleotide paired-end reads. Demultiplexing and generation of FASTQ files were performed using Illumina® CASAVA v1.8.2.

2.1.3. Data Generation, Summarization and Quality Control

For the BCCA samples, sequencing reads in FASTQ format were aligned to the reference human genome (hg19) using TopHat (v2.0.4) [132] with default parameters. The insert size mean and standard deviation were determined empirically based on alignment results using MISO [61]. Reads were realigned using TopHat and the insert size parameters. Alignment and quality metrics were calculated using RSeQC v2.3.3 [140]. To assess 3' bias per sample, the gene-body ratio metric was derived as the ratio between the average read coverage at 80% of the gene length and the average coverage at 20% of the gene length. Gene count estimates were derived using Python-based HTSeq-count [4] and the Ensembl v64 annotation in the General Transfer Format (GTF). Gene filtering was first conducted on normalized counts per million (CPM) calculated using edgeR and a modified version of the mixture model employed in the SCAN.UPC [95] Bioconductor package. A gene was

included in downstream analyses if the mixture model classified it as “on” (i.e. “signal”) in at least 15% of the samples.

In addition, to provide cross-platform compatibility, biomarker discovery and validations were performed on common 11,926 genes present on the RNA-Seq platform (Illumina HiSeq 2500 used with Ensembl v64 GTF) and two microarray platforms (Affymetrix GeneChip Human Gene 1.0 ST Array used with custom ENSG *Homo sapiens* CDF from Brainarray v11 and Affymetrix Human Genome U133A Array used with custom ENSG *Homo sapiens* CDF from Brainarray v16).

For the RPCI samples, gene counts were computed using RSEM [72] and Bowtie [69] with Ensembl v74 GTF annotation.

The sample and gene filtering methods for the (a) overlapping and independent bronchial brushing samples from subjects with and without COPD and PMLs profiled by microarray [122], (b) the bronchial brushing samples from subjects with and without lung cancer profiled by microarray [114,120], and (c) the TCGA LUSC tumor and adjacent normal biopsies from subjects with lung cancer profiled by RNA-Seq [137], are described in the corresponding publications.

2.1.4. Gene expression-based prediction of smoking status

Microarray data from Beane *et al*. [16] (Gene Expression Omnibus [GEO] accession ID GSE7895) was re-analyzed by Dr. Jennifer Beane using Robust Multi-array Average (RMA) [53] and the Ensembl Chip Definition File (CDF v16.0.0)

(http://brainarray.mbni.med.umich.edu/Brainarray/Database/CustomCDF/16.0.0/ensg.asp). The limma R package [99] was used to identify genes differentially expressed between current (n=52) and never (n=21) smokers as previously described [15]. Ninety-four genes (FDR<0.001) were differentially expressed between current and never smokers. The weighted voting algorithm [134] was trained on z-score normalized microarray data (n=73) across the 94 genes and used to predict smoking status in the 75 mRNA-Seq samples summarized with z-scored log2-transformed counts per million (log2 CPM).

2.1.5. Biomarker Discovery Pipeline

A gene expression biomarker discovery pipeline was developed to test thousands of parameter combinations to identify a biomarker capable of distinguishing between samples from subjects with and without PMLs. Samples were first assigned by batch (corresponding to sequencing lane) to either a discovery set (n=58) or a validation set (n=17), and the validation set was excluded from biomarker development. The biomarker was developed using subsets of the discovery set established by randomly splitting the samples into training (80%, n=46) and test (20%, n=12) sets 500 times, setting a common random seed of 150112. In each iteration, training set samples were passed through a series of six biomarker discovery steps, and the last step involving class prediction was in addition applied to the test set samples. The flowchart in Figure 2.1 and the following sections describe the pipeline steps in detail:

1. Balancing Signature

We tested gene signatures consisting either of an equal or unequal number of genes up- and down-regulated in subjects with dysplastic lesions.

2. Input Data Preprocessing

We tested three input data types. HTSeq-count (v0.5.4) [3] was used to derive gene count estimates (raw counts). In addition, Cufflinks (v2.0.2) [133] was used to derive reads per kilobase per million mapped reads (RPKM) using BAM files containing only properly paired reads. We also calculated log2-transformed counts per million (CPM) by applying edgeR (v3.8.6) [101] to raw counts using the "TMM" method (weighted trimmed mean of M-values [102]).

3. Gene Filtering

Signal-based gene filtering was conducted as described in detail in the Methods. In short, a gene was included in downstream analyses if the mixture model classified it as "on" in at least 1%, 5%, 10% or 15% of the samples. For CPM input data type, we recalculated CPM values using raw counts after filtering out genes.

4. Feature Selection

To identify genes differentially expressed (DE) between samples with and without premalignant lesions (PMLs), we applied several algorithms to our filtered dataset. The algorithms used were as follows:

1. edgeR: We applied the edgeR package (v3.8.6) [101] to raw counts only. After calculating normalization factors (calcNormFactors) and estimating common (estimateGLMCommonDisp) and tagwise (estimateGLMTagwiseDisp) dispersion factors, we identified DE genes associated with the presence of PMLs using a generalized linear model, correcting for sex, COPD status, and smoking status covariates. For balanced signatures, the sign of the log2-fold change of expression between conditions determined gene directionality. For all models regardless of balancing, gene importance was defined by FDR-adjusted p-value from likelihood ratio tests (glmLRT).
2. edgeRgb: We used the edgeR package as described in #1, additionally correcting for gb-ratio (described in section 2.1.3).
3. lm: We applied the limma package (v3.22.7) [134] to CPMs, RPKMs, or voom-transformed raw counts [70]. Voom transformation was applied using a linear model, adjusting for sex, COPD status, and smoking status covariates, after calculating normalization factors. We used the same model to identify DE genes associated with the presence of PMLs. For balanced signatures, the sign of the moderated t-statistic obtained via eBayes and topTable determined gene directionality. For all models regardless of balancing, gene importance was defined by the magnitude of the t-statistic.
4. lmgb: We used the limma package as described in #3, additionally correcting for gb-ratio (described in section 2.1.3).

5. glmnet: We applied the glmnet package (v1.9-8) [38] to CPMs, RPKMs, or voom-transformed raw counts (as in #3) to identify DE genes associated with the presence of PMLs. For balanced signatures, gene directionality was determined by the sign of the t-statistic obtained via limma by running a linear model described in #3. We carried out the following series of steps using all genes for unbalanced signatures and separately using up- and down-regulated genes for balanced signatures: First, RPKMs and CPMs were z-score normalized, while raw counts were voom-transformed. Then, due to the binary character of our response variable (dysplasia status), a logistic regression model was fit using the binomial distribution family and elastic net mixing parameter $\alpha = 0.5$ (indicating a tradeoff between ridge and lasso regressions). The standardize option was set to FALSE, causing the coefficients to be returned on the original scale, thus allowing their magnitude to be interpreted as gene importance. Next, a range of regularization parameters λ was generated via leave-one-out cross-validation (nfolds = 46), and the λ giving the minimum mean cross-validated error (lambda.min) was chosen to estimate the coefficients. Finally, DE genes were defined as having non-zero coefficients and then sorted by importance based on the coefficients' magnitude.
6. randomForest: We applied the randomForest package (v4.6-12) [73] to CPMs, RPKMs, and voom-transformed raw counts (as in #3), setting the number of trees (ntree) to 100 and importance to TRUE. For balanced signatures, the sign of the t-statistic as described in #5 determined gene directionality. For all models

regardless of balancing, gene importance was determined by the magnitude of the importance variable, defined as the mean decrease in accuracy over both conditions.

7. DESeq: We applied the DESeq package (v1.18.0) [2] to unmodified raw counts only. DE analysis to find genes associated with the presence of PMLs included data normalization (estimation of the effective library size), variance estimation, and inference for two experimental conditions, as outlined in the DESeq package vignette (https://www.bioconductor.org/packages/3.3/bioc/vignettes/DESeq/inst/doc/DESeq.pdf). For balanced signatures, the sign of the log2-fold change of expression between the two conditions determined gene directionality. For all models regardless of balancing, gene importance was defined by FDR.
8. SVA: We applied the sva package (v3.12.0) [71] to CPMs, RPKMs, or voom-transformed raw counts. Raw counts were voom-transformed using a linear model including only dysplasia status as the predictor variable. The number of surrogate variables (SVs) not associated with dysplasia status was estimated using the default approach of Buja and Eyuboglu [24] ("be" method). SVs were then identified using the empirical estimation of control probes ("irw" method), and up to 5 were added as covariates in the linear model (limma package). The adjusted model was then used to once again voom-transform raw counts, and subsequently fitted to identify DE genes associated with the presence of PMLs. For balanced

signatures, the sign of the moderated t-statistic obtained via topTable determined gene directionality. For all models regardless of balancing, gene importance was defined by the magnitude of the t-statistic.

9. pAUC (partial AUC) [80]: We applied the rowpAUCs function in the genefilter package (v1.48.1) [39] to CPMs, RPKMs, or voom-transformed raw counts (as in #3). We used 10 class label permutations and a sensitivity cutoff of 0.1 for a specificity range of 0.9-1. For balanced signatures, the sign of the moderated t-statistic obtained via limma's topTable determined gene directionality. For all models regardless of balancing, gene importance was defined by the magnitude of the t-statistic.

5. Gene Signature Size

After the feature selection step, we selected the top scoring 10, 20, 40, 60, 80, 100, or 200 genes, making sure that for balanced signatures, half originated from an ordered list of up-regulated genes, and half from an ordered list of down-regulated genes.

6. Prediction Method

For each set of genes, we applied multiple prediction methods to predict dysplasia status (presence of PMLs) in a training set of 46 samples and a test set of 12 samples. These training and test set samples differed in each iteration, which resulted from randomly splitting the 58 discovery set samples. The following prediction methods were used:

1. glmnet: We used glmnet (v1.9-8) [38] to first estimate a range of penalty parameters λ in 10-fold cross validation using the binomial distribution family parameter and α = 0 to ensure all feature-selected genes were included in predictions. Dysplasia status was then predicted as a binary class, using lambda.min penalty.
2. wv (weighted voting) [40]: We used the weighted voting algorithm to predict dysplasia status.
3. svm (Support Vector Machine) [30]: We used the svm function in the e1071 package (v1.6-7) [84] with linear kernel and 5-fold cross validation for class prediction.
4. rf (random forest): We used the randomForest package (v4.6-12) [73] with 1000 trees, requesting a matrix of class probabilities as output.
5. nb (Naïve Bayes): We used the naiveBayes function in the e1071 package (v1.6-7) with default parameters.

Each of the prediction algorithms generated a vector of predicted scores and a vector of predicted labels for all samples in the training and test sets.

Evaluating Performance

We considered all statistically and computationally viable combinations of the above six parameters. The predicted class labels calculated for each model (i.e., a combination of parameters), coupled with true class labels were then used to calculate performance metrics for the biomarker.

For each model, we calculated these metrics for each of the 500 iterations (different training and test sets assembled from the discovery set samples) and then averaged over all iterations. In addition to the standard performance metrics, we calculated model overfitting and gene selection consistency. The overfitting metric was calculated as the difference between the train set AUC and the test set AUC. Specifically, a model performing well on the training set but poorly on the test set would suggest a high degree of overfitting and thus achieve a high overfitting score. For each model, the gene selection consistency metric was calculated as the average ("normalized" to biomarker size in a given model) percentage of genes passing the gene filter, that were selected into the final gene committee in all 500 iterations:

$$consistency = 1 - \frac{\#\ unique\ genes\ in\ all\ iterations - biomarker\ size}{(biomarker\ size \times \#\ iterations) - biomarker\ size}$$

For example, a model resulting in a 10-gene biomarker would have the highest consistency (1) if it selected the same 10 genes in all 500 iterations (10 unique genes selected altogether). The same model would have the lowest consistency (0) if it selected a different set of 10 genes in all iterations (10 genes x 500 iterations = 5000 unique genes altogether).

Selecting Best Performing Model

In selecting the best model, we considered the degree of model overfitting, model gene selection consistency and test set AUC. First, we identified top 10% least overfitting models. Simultaneously, we identified top 10% most consistent models. Finally, the

model with the highest test set AUC among models fulfilling both criteria was chosen as the final model.

Selecting Final Gene Signature

Due to the nature of internal cross-validation, the biomarker genes selected by the best model may differ (only slightly, assuming the model is highly consistent) between iterations due to differences in the sample composition of the training set in each cross-validation. Therefore, we generated the final gene signature by training the best biomarker model using all 58 discovery set samples and best model parameters, allowing the pipeline to discover a new "consensus" signature.

Positivie and Negative Controls

The biomarker discovery pipeline was also used to develop control biomarkers. As positive controls, we used smoking status and sex phenotypes to identify biomarkers that could successfully distinguish former from current smokers, and females from males, respectively. As negative controls, we used randomly shuffled labels for dysplasia status, smoking status, and sex. Label shuffling was conducted preserving the association between gene expression profiles and remaining phenotypes; i.e., in the case of shuffled dysplasia status, only dysplasia status was shuffled while other phenotypes and the corresponding gene expression profile remained unchanged and linked to the same sample ID.

2.1.6. Biomarker Validation Pipeline

We tested the performance of the final biomarker using the biomarker discovery pipeline in validation mode. In this mode, the pipeline takes in the entire discovery set (n = 58) as the training set, and an external validation set as the test set. The test set is first corrected for gb-ratio (RNA-Seq quality metric) using limma, and the residual data is used as input. Both training and test sets are then z-score normalized. The pipeline is run using only the final model to generate prediction labels and prediction scores for the test set samples. Finally, pROC package (v1.8) [100] is used to visualize and quantify biomarker performance by plotting a ROC curve using prediction scores as the response and the dichotomous phenotype as the predictor, and extracting the AUC value from the resulting ROC object.

Detecting PML Presence Using Bronchial Brushings from Subjects With and Without PMLs Profiled by RNA-Seq

To validate the biomarker's ability to detect the presence of PMLs, we tested the performance of the biomarker in smokers with and without PMLs (n=17 samples) left out of the biomarker discovery process. To assess the robustness of the results, we randomly permuted dysplasia status labels 100 times, obtaining biomarker scores for all 17 samples in each of the iterations. We then concatenated the 100 newly generated biomarker score sets for randomized labels, creating a predictor vector consisting of 1700 scores. Similarly, we concatenated 100 identical copies of biomarker score sets for true labels,

creating a response vector of the same length. This allowed us to visualize the performance of the biomarker on true and randomized labels in a single ROC curve. Moreover, we evaluated the biomarker's platform dependence by testing its ability to detect PMLs in an overlapping and an independent set of microarray samples.

Detecting PML Progression Using Longitudinally-Collected Bronchial Brushings from Subjects With History of PMLs Profiled by RNA-Seq

To validate the biomarker's ability to predict sample progression/regression, we first used the biomarker to score the longitudinally collected RPCI samples (n=51). Next, we calculated the difference in scores between two consecutive time points for each patient (later time point biomarker score - earlier time point biomarker score). For example, a subject with 3 samples from 3 different time points would have 3 scores, and thus two score differences could be calculated; a subject with 2 samples from 2 time points would have 2 scores, and thus 1 score difference. Each pair of samples was assigned a "progressing/stable" or "regressing" phenotype. A "progressing/stable" phenotype indicated that the worst histological grade of PMLs sampled during the baseline procedure increased in severity or remained unchanged at follow-up; while a "regressing" phenotype indicated that the worst histological grade of PMLs sampled at baseline decreased in severity at follow-up. We quantified the ability of the score difference to predict the "progression/regression" phenotype by plotting a ROC curve,

using the vector of score differences as the predictor variable, and the progression/regression phenotype as the response variable.

Detecting Lung Cancer Presence Using Bronchial Brushings from Subjects With and Without Lung Cancer Profiled by Microarray

To validate the biomarker's ability to detect the presence of lung cancer, we tested the model in microarray and RNA-Seq samples from subjects with and without lung cancer.

Functional Enrichment

The final biomarker genes' role in relevant biological pathways and processes was evaluated using Enrichr [28]. Enrichment was tested using Fisher's Exact Test in gene enrichment categories such as BioCarta, Reactome, GO, KEGG, etc, calculated for the gene overlap between the biomarker genes and genes implicated in each considered pathway.

2.2. Results

2.2.1. Sample Population

Cytologically normal epithelial cells were collected via bronchial airway brushings using autofluorescence bronchoscopy from current and former smokers at the BCCA. Sample and gene filtering yielded 13,870 out of 51,979 genes and 82 samples (n=2 excluded due to quality or sex annotation mismatches) for analysis. Data from

Beane *et al.* [16] was used to predict the smoking status of the 82 samples. Airway brushings with an assigned histological grade of metaplasia (n=7), were removed from further analysis due to classification uncertainties. The remaining 75 samples were dichotomized into two groups: samples with no evidence of PMLs (brushes from subjects with no abnormal fluorescing areas or those corresponding to biopsies having normal or hyperplasia histology, n=25); and samples with evidence of PMLs (brushes corresponding to biopsies having mild, moderate, or severe dysplasia, n=50).

Important clinical covariates including COPD status, self-reported smoking history and genomically-derived smoking status were not significantly different between the two groups (Table 2.2). For biomarker development, the 75 BCCA samples were split by batch (sequencing lane) and used separately in biomarker discovery (n=58) and validation (n=17) (Table 2.2)

The change in biomarker score as a predictor of PML progression was tested in 51 RPCI RNA-Seq samples (Table 2.3 and Table 2.4).

The biomarker's ability to detect PMLs was further validated in 36 overlapping and 158 independent samples from current and former smokers enrolled in chemoprevention studies [122,123] (Table 2.5 and Table 2.6), in addition enabling the concurrent evaluation of the biomarker's performance on a microarray platform.

Moreover, the biomarker's ability to distinguish bronchial brushing samples from subjects with and without lung cancer was tested in two microarray datasets (n=164 and n=299) [114,120](Table 2.7 and Table 2.8).

Finally, the biomarker's ability to distinguish tumor samples from adjacent normal samples originating from subjects with squamous cell lung cancer, was evaluated using 223 TCGA RNA-Seq LUSC samples [137] (Table 2.9).

2.2.2. Performance Metrics

A total of 7,700 parameter combinations (models) were tested using the biomarker discovery pipeline. The validity of the pipeline was evaluated by examining performance metrics across all models and all 500 iterations for dysplasia status as well as positive and negative controls.

Overall, the performance of the biomarker on dysplasia status as summarized by the average test set AUC across 500 iterations ranged between 40% and 80%, with a median of 69% and standard deviation of 5%. The train set AUC average ranged between 53% and 100%, with a median of 97% and standard deviation of 6%. The overall performance of the biomarker on positive and negative control has been summarized in (Figure 2.3).

The performance of the biomarker was also summarized by the individual parameters of each model. The test and train set AUCs were evaluated separately by balancing, data type, gene filter, etc. These metrics reflect the performance of one parameter summarized over there remaining parameters, e.g. Figure 2.2 illustrates that 200-gene biomarkers performed better than smaller size ones, regardless of balancing,

data type, gene filter, etc., and that RPKMs gave slightly better results that CPM or raw counts regardless of other parameters.

Model overfitting defined as the difference between test and train set AUCs was summarized for dysplasia status and positive and negative controls along with model consistency in Figure 2.4.

All performance metrics including average sensitivity, specificity, PPV and NPV were summarized in Table 2.10.

2.2.3. Selection of Best Model and Final Gene Signature

In selecting the final model, we first considered choosing the model with the highest test set AUC. However, among the 7,700 models we tested, 27 had an AUC > 0.79 and 147 had an AUC > 0.78, which is within 0.02 of the maximum AUC observed for all models (max AUC = 0.80). Thus, the decision to simply pick the top scoring model would be arbitrary, as almost all representatives of parameters considered were among the top 147 scoring models. To circumvent this limitation, in deciding about the final model we took into consideration the degree of model overfitting and consistency. We picked the model with the highest test set AUC from a subset of models that ranked among the top 10% (770 models) of least overfitting models and the top 10% (770 models) of most consistent models.

The final model had test set AUC of 0.78 (ranking 216th among best test set AUC models), an 11% overfitting score (ranking 57th among least overfitting models), and a

98% consistency score (ranking 446th among most consistent models) in cross-validation (Table 2.11). The selected model produced an imbalanced signature of 200 features, after filtering out genes expressed in less than 15% of samples and applying a quality-corrected linear model to CPM for feature selection and weighted voting algorithm for classification.

Although the final chosen model did not achieve the highest test set AUC, incorporating other performance metrics allowed us to select a model that better captures the signal without overfitting, producing more robust results.

2.2.4. Positive and Negative Controls

The biomarker's ability to distinguish samples from subjects with and without PMLs was evaluated against its power to tell current smokers from former, and males from females. Many tested models performed very well on the positive controls, which confirms the ability of the pipeline to find relevant biomarkers for traits with especially strong signal, such as smoking status and sex. As expected, the biomarker performed poorly on all three negative controls defined by shuffling the phenotype labels for dysplasia, smoking and sex, resulting in random class assignments and AUCs of around 0.5 (Figure 2.3).

2.2.5. Validations

Detecting Presence of PMLs

Validation set (n=17). The biomarker's ability to distinguish samples from subjects with and without PMLs was first evaluated in the validation set (n=17) of samples left out of the biomarker discovery process. The model achieved AUC of 0.75, and correctly predicted 5/5 of dysplasias (sens = 100%) and 9/12 of normals (spec = 75%). To assess the biomarker's robustness, the dysplasia status variable was shuffled 100 times. The model achieved AUC of 51% in predicting shuffled dysplasia status (negative control for dysplasia status), confirming the biomarker's expected poor performance predicting random signal (Figure 2.5).

Overlapping microarray PML set (n=36). A total of 238 samples collected from subjects with and without COPD were profiled on Affymetrix Human Genome U133A Array [122]. A subset of the original BCCI samples (n=36) originated from patients enrolled in this study and were subjected to profiling by both microarray and RNA-Seq. These additional microarray samples were used as an overlapping validation set (Table 2.5). The model achieved AUC of 73%, and correctly predicted 10/11 of dysplasias (sens = 91%) and 13/25 of normals (spec = 52%) (Figure 2.6)

Independent microarray PML set (n=158). A remainder of 158 samples analyzed in the same paper [122] and not profiled by RNA-Seq, were used as an independent

validation set (Table 2.6). The model achieved AUC of 64%, and correctly predicted 24/35 of dysplasias (sens = 68%) and 51/123 of normals (spec = 42%) (Figure 2.7)

Detecting Progression of PMLs

RNA-Seq longitudinal PML set (n=51). To validate the biomarker's ability to predict sample progression/regression, we used longitudinally collected RPCI samples. Among the 26 subjects, 18 had brushings performed at two time points, and five at three time points (Table 2.3). Subjects with a single time point (n=3) were removed from further analysis. First, we examined the performance of the biomarker scores in identifying samples from subjects with and without PMLs. When Metaplasias were classified as Normals, the model achieved an AUC of 62.3%, correctly predicting 21/31 samples from subjects with PMLs (sens =68%) and 11/20 samples from subjects with no evidence of PMLs (spec = 55%) (Figure 2.8).

We sought to find out if the differences in biomarker scores between consecutive time points would improve the performance. To test this hypothesis, we analyzed 28 score differences, and samples with a negative score difference were designated as regressing, and those with a positive score as progressing or stable. We quantified the ability of the score difference to predict the progression/regression phenotype by plotting a ROC curve and extracting the AUC as described in Methods. The model achieved an AUC of 74.9% (Figure 2.9).

Detecting Presence of Lung Cancer

Lung cancer bronchial brushes microarray set 1 (n=164). The biomarker's ability to distinguish samples from subjects with and without lung cancer was evaluated in additional bronchial brushings profiled on microarray [120]. Cancer samples were assigned the dysplasia phenotype. The model achieved 69.3% AUC , and correctly predicted 57/78 cancer samples (sens = 73%) and 48/86 normals (spec =56%). (Figure 2.10).

Lung cancer bronchial brushes microarray set 2 (n=299). Additional bronchial samples profiled on microarray as part of the AEGIS trial [114,141]were used as a validation set, and similarly, cancers were treated as dysplasias. The model achieved AUC of 52.3%, and correctly predicted 137/223 cancer samples (sens = 61%) and 32/76 normals (spec = 42%) (Figure 2.11).

TCGA Lung SCC tumor RNA-Seq set (n=553). Finally, 553 tumor and adjacent normal samples from subjects with lung squamous cell carcinoma catalogued in the LUSC dataset by the Cancer Genome Atlas Data Portal [137], were assessed within the biomarker framework. Tumor samples were assigned the status of dysplasia. The model achieved AUC of 81.8%, correctly predicting 315/502 (sens=63%) tumor samples and 43/51 adjacent normal (spec=84%) (Figure 2.12).

2.2.6. Biological Enrichment and Pathway Analysis

The 200 final biomarker genes were evaluated for their potential role in biological pathways using Enrichr. Pathways involved in respiratory electron transfer chain,

formation of ATP, oxidative phosphorylation and metabolism were strongly enriched among genes up-regulated in the airways of subjects with PMLs. Other up-regulated pathways included p53 signaling, and mitochondrial biogenesis. Among upregulated transcription factors, we observed the activating transcription factor 2 (*ATF2*) and the estrogen receptor 1 (*ESR1*). Down-regulated pathways included the *Pelp1* and *CARM1* modulation of estrogen receptor.

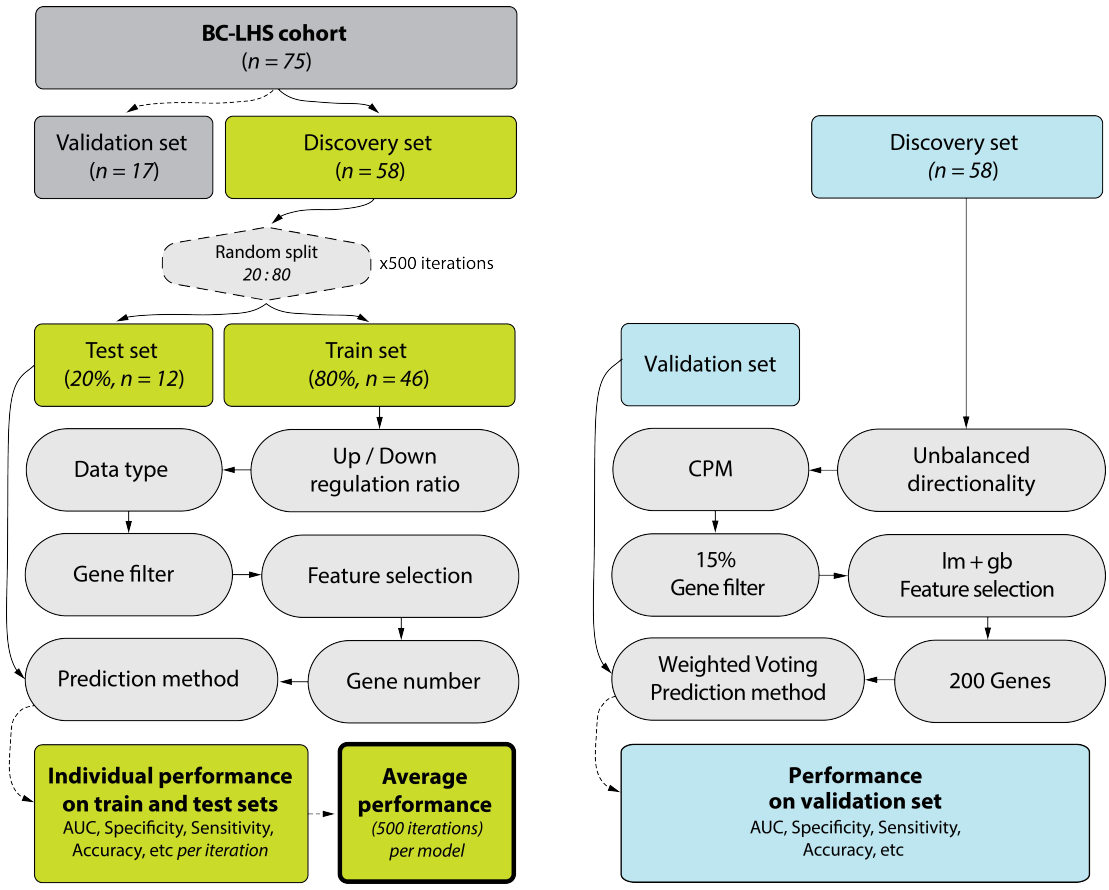


Figure 2.1 Biomarker discovery and validation pipelines.
(**LEFT**) Samples (n=75) were split into a discovery set (n=58) and a validation set (n=17). The pipeline was run 500 times, and each time the discovery set was randomly split into training (80% of samples, n=46) and test (20% of samples, n=12) sets. The training set samples were used to train the biomarker using all statistically and computationally feasible combinations of pipeline parameters, including: 1. Up- / down-regulation ratio: TRUE or FALSE (see Balancing signature). 2. Data type: raw counts, RPKM or CPM (see Input data preprocessing). 3. Gene filter: genes with signal in at least 1%, 5%, 10%, or 15% of samples (see Gene filter). 4. Feature selection: edgeR, edgeR correcting for gb-ratio, limma, limma correcting for gb-ratio, glmnet, random forest, DESeq, SVA, or partial AUC (see Feature selection). 5. Gene number: 10, 20, 40, 60, 80, 100, or 200 genes (see Biomarker size). 6. Prediction method: weighted voting, random forest, SVM, naïve bayes, or glmnet (see Prediction method). (**RIGHT**) In validation mode, the pipeline is run only once. The entire discovery set (n=58) is used to train the biomarker with parameters selected for the final model. The selected prediction algorithm is then applied to the validation set and performance metrics are calculated as before.

Table 2.1 Performance measures used to evaluate performance of the biomarker.
TP = true positives; FP = false positives; TN = true negatives; FN = false negatives; MCC = Matthews's Correlation Coefficient; and AUC = Area Under the Curve.

AUC for ROC (Receiver Operating Characteristic)	
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$
Sensitivity	$\frac{TP}{TP + FN}$
Specificity	$\frac{TN}{FP + TN}$
Positive Predictive Value (PPV)	$\frac{TP}{TP + FP}$
Negative Predictive Value (NPV)	$\frac{TN}{TN + FN}$
Matthew's Correlation Coefficient (MCC)	$\frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$
MAQCII	$0.5 \times AUC + 0.25 \times (MCC + 1)$

Table 2.2 Demographic characteristics of the discovery and validation sets stratified by dysplasia status.
Data are means (SD) for continuous variables and proportions with percentages for dichotomous variables. P-values are for the comparison of all-Caucasian subjects with and without premalignant lesions (dysplasia), using two sample t-tests for continuous variables or Fisher's exact test for categorical variables.

	Discovery Set				Validation Set			
Factor	**Overall**	**Normal**	**Dysplasia**	**p-value**	**Overall**	**Normal**	**Dysplasia**	**p-value**
No. Samples	58	20	38		17	5	12	
Age	62.7 (7.1)	64.1 (5.8)	61.9 (7.6)	0.24	63.9 (8.6)	66 (5.8)	63 (9.7)	0.45
Male	37/58 (63.8)	12/20 (60)	25/38 (65.8)	0.78	14/17 (82.4)	4/5 (80)	10/12 (83.3)	1
Current smoker	28/58 (48.3)	9/20 (45)	19/38 (50)	0.79	8/17 (47.1)	2/5 (40)	6/12 (50)	1
Pack years	48.2 (16.9)	49.4 (18.9)	47.5 (15.9)	0.71	44.6 (12.9)	40.5 (11.6)	46.3 (13.5)	0.39
FEV1% Predicted	86.5 (17.7)	87.8 (16.7)	85.7 (18.5)	0.66	69.5 (16.2)	71 (17.7)	68.9 (16.3)	0.83
FEV1/FVC Ratio	72.1 (7.7)	75.1 (6.3)	70.4 (8)	*0.02*	67 (8.1)	66.8 (8.5)	67.1 (8.3)	0.95
COPD (FEV1%<80 & FEV1/FVC<70)	11/58 (19)	2/20 (10)	9/38 (23.7)	0.3	11/17 (64.7)	3/5 (60)	8/12 (66.7)	1
Histology				*<0.001*				*<0.001*
Normal	11/58 (19)	11/20 (55)			1/17 (5.9)	1/5 (20)		
Hyperplasia	9/58 (15.5)	9/20 (45)			4/17 (23.5)	4/5 (80)		
Metaplasia	0/58 (0)				0/17 (0)			
Mild Dysplasia	29/58 (50)		29/38 (76.3)		6/17 (35.3)		6/12 (50)	
Moderate Dysplasia	6/58 (10.3)		6/38 (15.8)		6/17 (35.3)		6/12 (50)	
Severe Dysplasia	3/58 (5.2)		3/38 (7.9)		0/17 (0)		0/12 (0)	

Table 2.3 Demographic characteristics of the RNA-Seq cross-sectional bronchial brushing dataset stratified by dysplasia status.
Data are means (SD) for continuous variables and counts for dichotomous variables. Top table summarizes data by samples, and bottom table summarizes data by subjects. P-values are for the comparison between the Dysplasia and Normal groups, using two sample t-tests for continuous variables or Fisher's exact test for categorical variables.

Factor	Overall	Dysplasia	Normal	p-value
No. Samples	51	31	20	
Age	58.2 (6.6)	58.52 (6.93)	57.62 (6.15)	0.637
Sex				0.567
female	27	15	12	
male	24	16	8	
Race				*< 0.001*
Caucasian	49	29	20	
Other	2	2	0	
Lung Cancer	0	0	0	
Smoking				
current	25	18	7	
former	25	12	13	
never	1	1	0	
Pack Years	48.6 (22.58)	46.58 (21.42)	51.7 (24.5)	0.435
Worst Histological Lesion Observed				*< 0.001*
Normal	5/51 (9.8)		5/51 (9.8)	
Hyperplasia	6/51 (11.8)		6/51 (11.8)	
Metaplasia	9/51 (17.6)		9/51 (17.6)	
Mild Dysplasia	3/51 (5.9)	3/51 (5.9)		
Moderate Dysplasia	20/51 (39.2)	20/51 (39.2)		
Severe Dysplasia	8/51 (15.7)	8/51 (15.7)		

Factor	Overall
No. of subjects	23
No. of procedures per subject (mean [range])	2.2 (2-3 visits)
Days between visits (mean [range])	344 (98-721 days)
Age	58.26 (6.83)
Sex	
female	12
male	11
Race	
Caucasian	22
African American	1
Lung Cancer	0
Smoking	
current	10
former	10
former	3
Pack Years	49.22 (23.74)

Table 2.4 Demographic characteristics of the RNA-Seq paired PML dataset stratified by PML progression/regression.

Data are means (SD) for continuous variables and counts for dichotomous variables. P-values are for the comparison of sample pairs from subjects with regressing and progressing/stable premalignant lesions, using two sample t-tests for continuous variables or Fisher's exact test for categorical variables

Factor	Overall	Regressing	Progressing Stable	p-value
No. Samples	51	34	22	
No. Sample Pairs	28	17	11	
No. Patients**	23	16*	10*	
Age	58.1 (6.5)	58.4 (6.9)	57.6 (6.1)	1
Sex				0.7
female	15	10	5	
male	13	7	6	
Race				
Caucasian	28	17	11	
Lung Cancer	0	0	0	
Smoking				0.148
ever	27	17	10	
never	1	0	1	
Pack Years	48.1 (22)	49.8 (24.8)	45.4 (17.6)	1
Histological Grade Change	-0.9 (1.7)	-1.9 (1.0)	0.7 (1.3)	*<0.001*
Time between Procedures (Days)	343.8 (171.9)	350.9 (199.6)	332.8 (125.9)	0.77

Table 2.5 Demographic characteristics of the microarray overlapping PML dataset stratified by dysplasia status.
Data are means (SD) for continuous variables and counts for dichotomous variables. P-values are for the comparison of subjects with and without premalignant lesions (dysplasia), using two sample t-tests for continuous variables or Fisher's exact test for categorical variables.

	Overall	Dysplasia	Normal	p-value
No. Samples	36	11	25	
Sex				0.159
female	16	7	9	
male	20	4	16	
Age	65.63 (5.86)	68.16 (5.44)	64.52 (5.8)	0.086
COPD				0.224
no	26	6	20	
yes	10	5	5	
Smoking				1
current	17	5	12	
former	19	6	13	
Pack Years	46.05 (15.15)	47.59 (5.005)	45.41 (17.84)	0.708
Histology				*< 0.001*
Normal	12	0	12	
Hyperplasia	13	0	13	
Metaplasia	0	0	0	
MildD	9	9	0	
ModD	2	2	0	
SevD	0	0	0	

Table 2.6 Demographic characteristics of the microarray independent PML dataset stratified by dysplasia status.
Data are means (SD) for continuous variables and counts for dichotomous variables. P-values are for the comparison of subjects with and without premalignant lesions (dysplasia), using two sample t-tests for continuous variables or Fisher's exact test for categorical variables.

	Overall	Dysplasia	Normal	p-value
No. Samples	158	35	123	
Sex				0.052
female	64	9	55	
male	94	26	68	
Age	64.44 (5.44)	65.05 (5.11)	64.26 (5.54)	0.454
COPD				*0.048*
no	97	16	81	
yes	61	19	42	
Smoking				1
current	59	13	46	
former	99	22	77	
Pack Years	48.05 (20.99)	48.63 (16.6)	47.9 (22.07)	0.863
Histology				*< 0.001*
Normal	55	0	55	
Hyperplasia	68	0	68	
Metaplasia	0	0	0	
MildD	23	23	0	
ModD	9	9	0	
SevD	3	3	0	

Table 2.7 Demographic characteristics of the microarray lung cancer bronchial brushing dataset 1 stratified by cancer status.
Data are means (SD) for continuous variables and counts for dichotomous variables. P-values are for the comparison of subjects with and without lung cancer, using two sample t-tests for continuous variables or Fisher's exact test for categorical variables.

	Overall	Cancer	Normal	p-value
No. Samples	164	78	86	
Sex				0.592
female	41	18	23	
male	122	60	62	
NA	1	0	1	
Age	58.15 (14.32)	64.54 (9.63)	52.28 (15.42)	*< 0.001*
NA	1	0	1	
Smoking				0.654
current	130	60	70	
former	33	18	15	
NA	1	0	1	
Race				*< 0.001*
Caucasian	110	67	43	
Other	53	11	42	
NA	1	0	1	
Pack Years	44.92 (31.95)	54.92 (26.77)	35.73 (33.67)	*< 0.001*
NA	1			

Table 2.8 Demographic characteristics of the microarray lung cancer bronchial brushing dataset 2 stratified by cancer status.
Data are means (SD) for continuous variables and counts for dichotomous variables. P-values are for the comparison of subjects with and without lung cancer, using two sample t-tests for continuous variables or Fisher's exact test for categorical variables.

	Overall	Cancer	Normal	p-value
No. Samples	299	223	76	
Sex				0.178
female	123	97	26	
male	176	126	50	
Age	62.27 (12.1)	64.48 (10.55)	55.8 (13.98)	*< 0.001*
COPD				*0.0315*
no	191	139	52	
unknown	4	1	3	
yes	104	83	21	
Smoking				0.107
current	127	101	26	
former	172	122	50	
Pack Years	43.79 (30.54)	47.84 (31.3)	31.64 (24.57)	< 0.001

Table 2.9 Demographic characteristics of the RNA-Seq lung tumor biopsy dataset 1 stratified by cancer status.
Normal refers to biopsy taken from area adjacent to the tumor. Data are means (SD) for continuous variables and counts for dichotomous variables. P-values are for the comparison of subjects with and without lung cancer, using two sample t-tests for continuous variables or Fisher's exact test for categorical variables.

	Overall	Cancer	Normal	p-value
No. Samples	553	502	51	
Sex				0.879
female	144	130	14	
male	408	371	37	
Age	67.3 (8.59)	67.2 (8.58)	68.24 (8.65)	0.414
Smoking				0.654
current	213	195	18	
former	340	307	33	
Pack Years	52.68 (31.01)	52.69 (31.03)	52.57 (31.21)	0.98

Table 2.10 Biomarker performance in cross-validation.
Summary of performance metrics for 7700 considered models. Each performance measure represents summaries over 500 iterations per model. The first row summarizes parameters used by the 7700 models. The second and third rows summarize performance for train and test set, respectively. ACC=accuracy, SENS=sensitivity, SPEC=specificity, PPV=positive predictive value, NPV=negative predictive value, MCC=Mathew's Correlation Coefficient, AUC=Area Under the (ROC) Curve, MAQC2=performance metric derived in MicroArray Quality Control Study (Table 2.1).The fourth row summarizes model overfitting and consistency.

transformation	sample_filter	data_type	gene_filter	feature_selection	number_features	prediction_mode	balanced
raw:7700	bal :3850	counts:3500	0.01 :1540	glmnet :1050	Min. : 10.0	glmnet:1540	Mode :logical
	imbal:3850	cpm :2100	0.05 :1540	lm :1050	1st Qu.: 20.0	nb :1540	FALSE:3850
		rpkm :2100	0.1 :1540	lmgb :1050	Median : 60.0	rf :1540	TRUE :3850
			0.15 :1540	pauc :1050	Mean : 72.9	svm :1540	
			lam.grt.cor:1540	rf :1050	3rd Qu.:100.0	wv :1540	
					Max. :200.0		

ACC.tr.dysp	SENS.tr.dysp	SPEC.tr.dysp	PPV.tr.dysp	NPV.tr.dysp	MCC.tr.dysp	AUC.tr.dysp	MAQC2.tr.dysp
Min. :0.621	Min. :0.117	Min. :0.579	Min. :0.325	Min. :0.688	Min. :0.087	Min. :0.525	Min. :0.539
1st Qu.:0.817	1st Qu.:0.820	1st Qu.:0.816	1st Qu.:0.715	1st Qu.:0.903	1st Qu.:0.629	1st Qu.:0.911	1st Qu.:0.860
Median :0.909	Median :0.895	Median :0.965	Median :0.921	Median :0.944	Median :0.797	Median :0.969	Median :0.933
Mean :0.892	Mean :0.864	Mean :0.907	Mean :0.854	Mean :0.934	Mean :0.772	Mean :0.947	Mean :0.917
3rd Qu.:0.987	3rd Qu.:0.985	3rd Qu.:0.996	3rd Qu.:0.991	3rd Qu.:0.992	3rd Qu.:0.971	3rd Qu.:0.999	3rd Qu.:0.992
Max. :1.000	Max. :1.000	Max. :1.000	Max. :1.000	Max. :1.000	Max. :1.000	Max. :1.000	Max. :1.000

ACC.ts.dysp	SENS.ts.dysp	SPEC.ts.dysp	PPV.ts.dysp	NPV.ts.dysp	MCC.ts.dysp	AUC.ts.dysp	MAQC2.ts.dysp
Min. :0.481	Min. :0.041	Min. :0.527	Min. :0.167	Min. :0.596	Min. :-0.105	Min. :0.403	Min. :0.429
1st Qu.:0.629	1st Qu.:0.395	1st Qu.:0.685	1st Qu.:0.502	1st Qu.:0.702	1st Qu.: 0.194	1st Qu.:0.660	1st Qu.:0.627
Median :0.655	Median :0.468	Median :0.785	Median :0.549	Median :0.727	Median : 0.255	Median :0.692	Median :0.659
Mean :0.654	Mean :0.476	Mean :0.756	Mean :0.550	Mean :0.729	Mean : 0.248	Mean :0.693	Mean :0.658
3rd Qu.:0.683	3rd Qu.:0.594	3rd Qu.:0.815	3rd Qu.:0.601	3rd Qu.:0.758	3rd Qu.: 0.319	3rd Qu.:0.732	3rd Qu.:0.695
Max. :0.747	Max. :0.729	Max. :0.979	Max. :0.747	Max. :0.831	Max. : 0.464	Max. :0.799	Max. :0.763

overfitting.dysp	consistency.dysp
Min. :0.0431	Min. :0.23
1st Qu.:0.2148	1st Qu.:0.94
Median :0.2593	Median :0.96
Mean :0.2545	Mean :0.94
3rd Qu.:0.2982	3rd Qu.:0.97
Max. :0.4844	Max. :0.99
	NA's :225

Table 2.11 Best model performance

Overview of parameters and performance metrics for the selected model. The first row summarizes parameters used by the best model. The second and third rows show performance of the best model on train and test set, respectively. ACC=accuracy, SENS=sensitivity, SPEC=specificity, PPV=positive predictive value, NPV=negative predictive value, MCC=Mathew's Correlation Coefficient, AUC=Area Under the (ROC) Curve, MAQC2=performance metric derived in MicroArray Quality Control Study (Table 2.1).The fourth row shows the best model overfitting and consistency.

sample_filter	data_type	gene_filter	feature_selection	number_features	prediction_mode	balanced	phenotype.dysp
imbal	cpm	0.15	lmgb	200	wv	FALSE	dysplasia_status

ACC.tr.dysp	SENS.tr.dysp	SPEC.tr.dysp	PPV.tr.dysp	NPV.tr.dysp	MCC.tr.dysp	AUC.tr.dysp	MAQC2.tr.dysp
0.8135	0.8792	0.7795	0.6754	0.9263	0.6296	0.8862	0.8505

ACC.ts.dysp	SENS.ts.dysp	SPEC.ts.dysp	PPV.ts.dysp	NPV.ts.dysp	MCC.ts.dysp	AUC.ts.dysp	MAQC2.ts.dysp
0.6913	0.705	0.6834	0.576	0.8205	0.3914	0.777	0.7364

overfitting.dysp	consistency.dysp
0.1092	0.9814

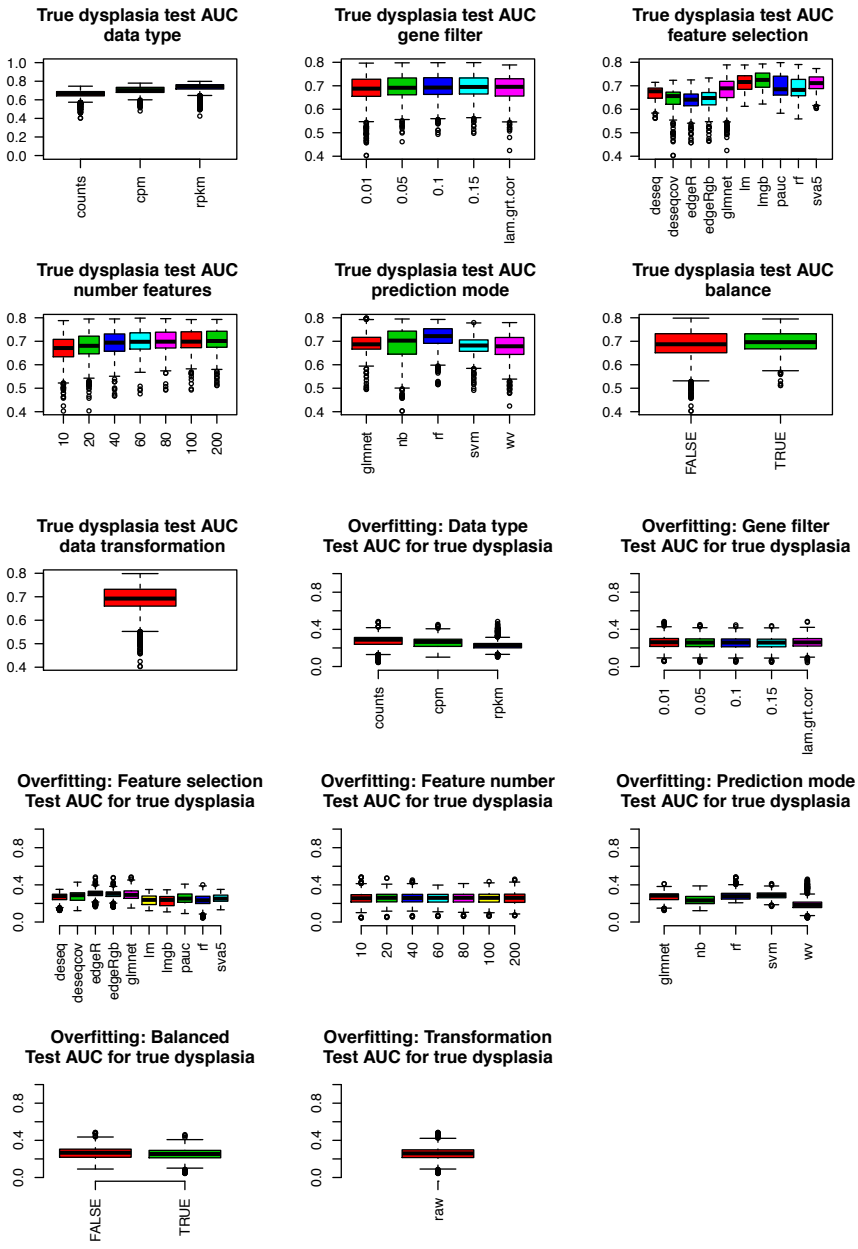


Figure 2.2 Biomarker performance in cross validation.
(1-7 row-wise) Boxplots summarizing AUC on test set across 7700 models stratified by model parameter. (8-14 row-wise) Boxplots summarizing degree of overfitting across 7700 models stratified by model parameters.

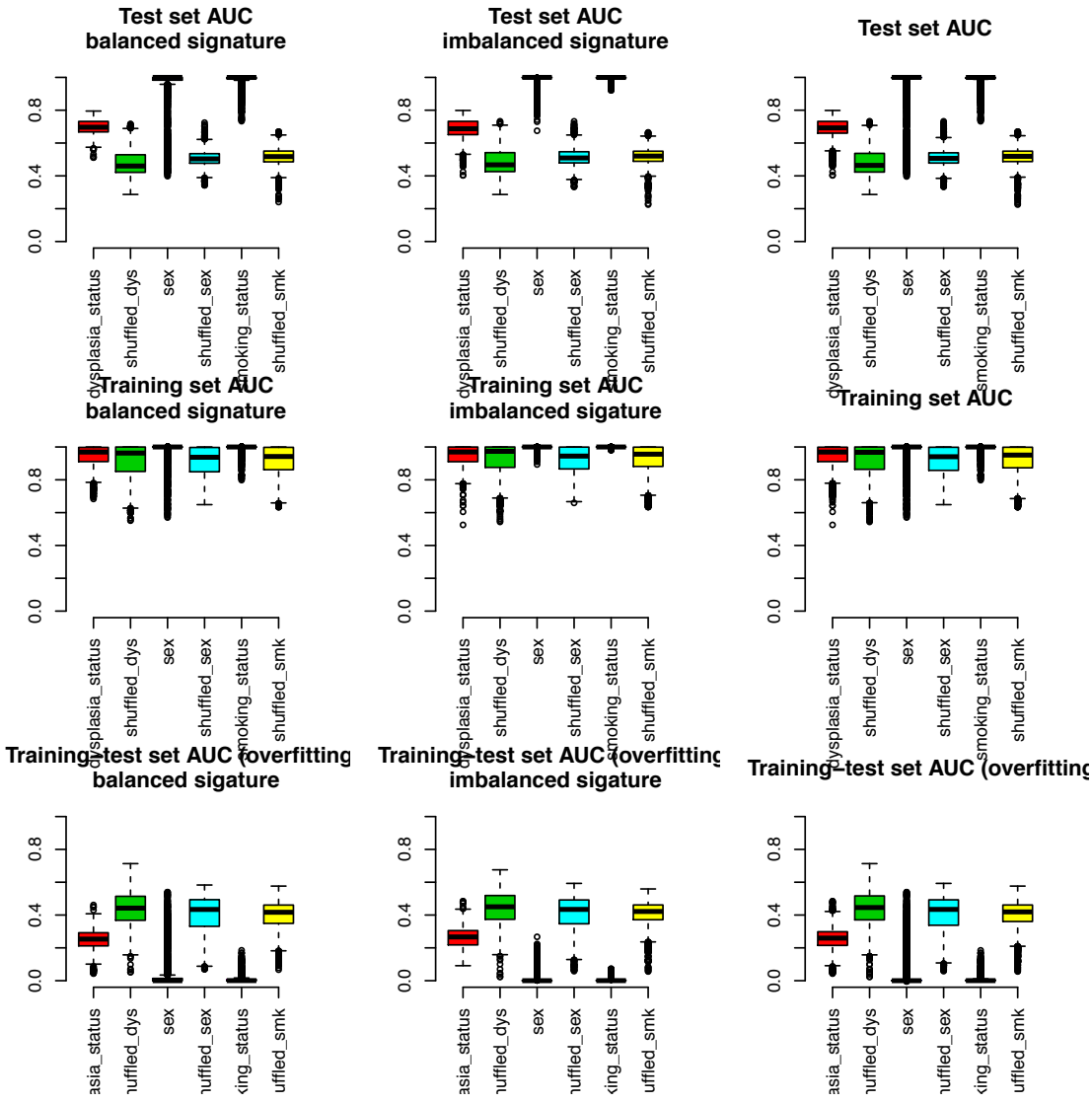


Figure 2.3 Biomarker performance on positive and negative controls.
Summaries of performance of 7700 models predicting PML presence, sex, and smoking status, as well as shuffled equivalents (negative controls).Summary of AUC in test samples (first row), AUC in training samples (second row), and overfitting (third row) stratified by balancing.

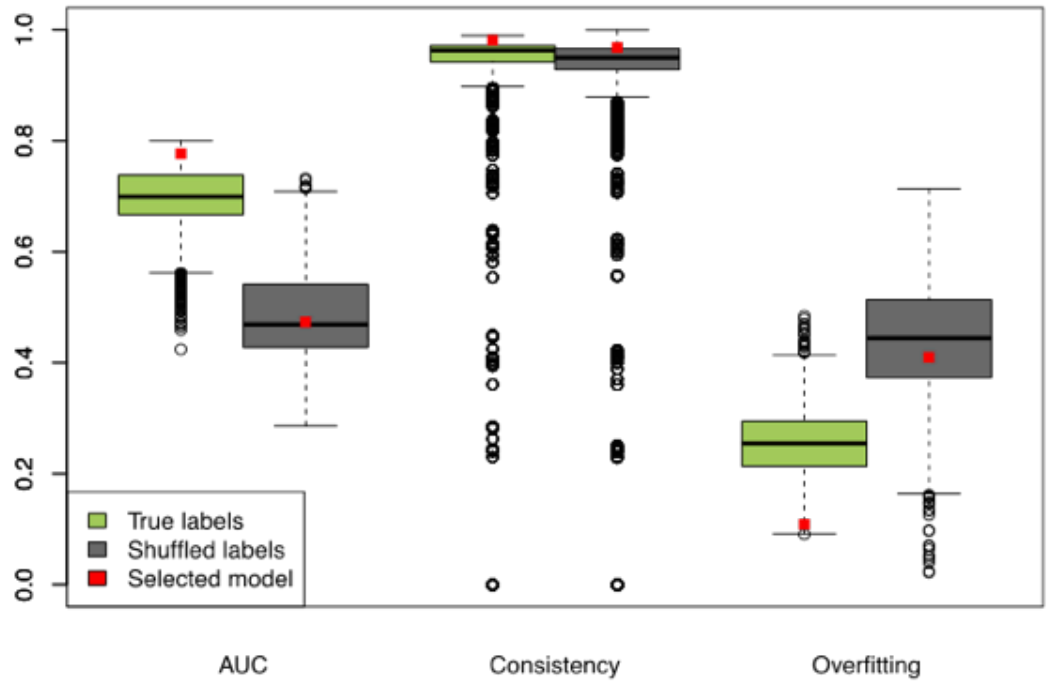


Figure 2.4 Summary of biomarker performance measures used to select the best model. 7,700 models for PML presence prediction were tested on true and shuffled dysplasia labels and their performance was summarized across 500 cross-validations using the area under the curve (AUC). Model consistency was defined as model's ability to select the same genes in all 500 cross-validations. Model overfitting was defined as the difference between train set AUC and test set AUC. The final model was chosen as one having the highest possible test set AUC among highly consistent and least overfitting models, and is highlighted in red.

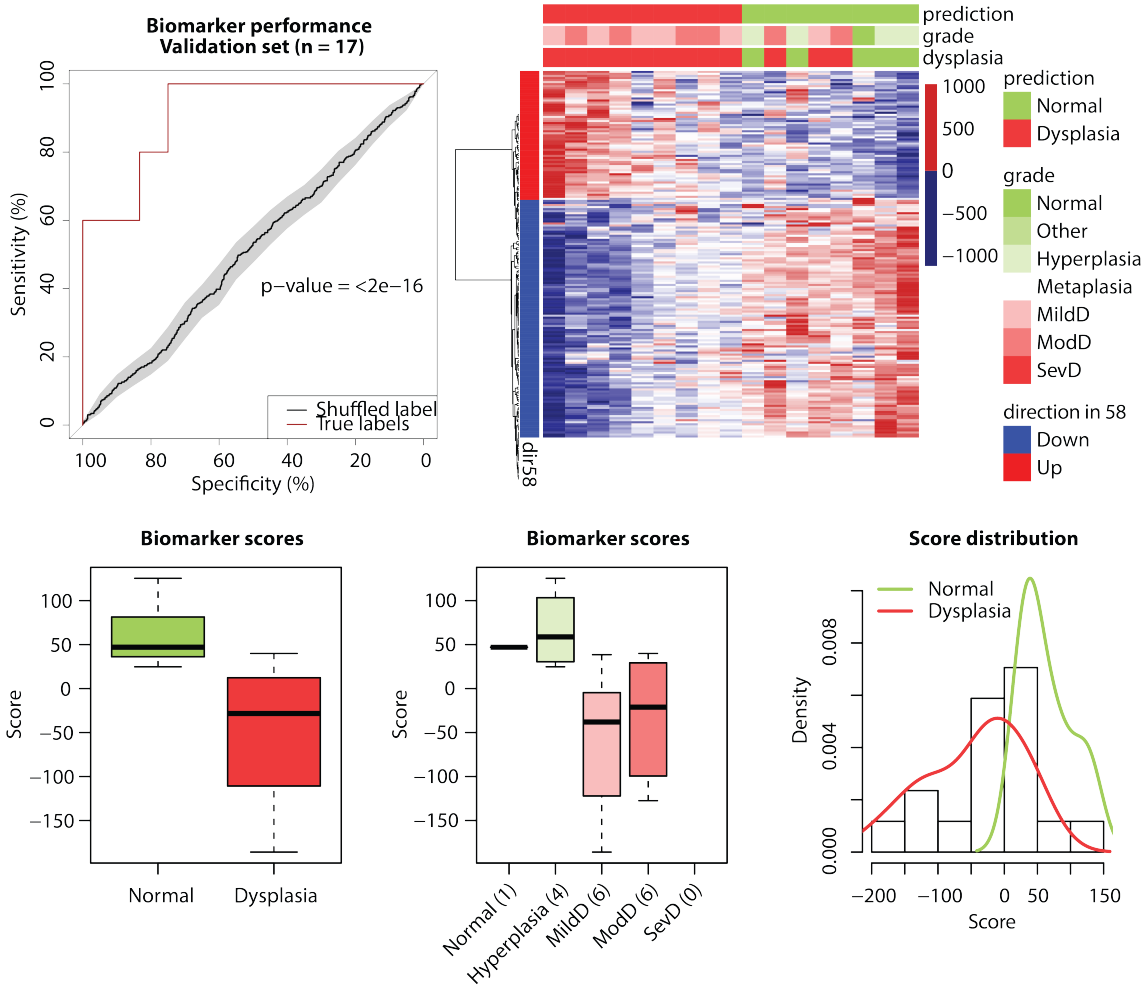


Figure 2.5 Biomarker performance in validation set (n=17).
To evaluate biomarker's performance predicting PML presence, the biomarker was trained on discovery set samples (n=58) and tested in validation set samples (n=17) left out of the discovery process. (**TOP LEFT**) ROC curve summarizing performance in the validation set (red) and in a negative control set constructed by shuffling the dysplasia labels in the validation samples 100 times (black with grey confidence interval area). (**TOP RIGHT**) Heatmap showing 200 biomarker genes (rows) and 17 validation samples (columns). Horizontal color bars correspond to dysplasia / normal phenotype: top bar shows biomarker prediction, and middle and bottom bars show actual dysplasia grade and status. Vertical color bar corresponds to the directionality of genes in 58 discovery samples. (**BOTTOM LEFT**) Boxplot showing biomarker score stratified by dysplasia status. (**BOTTOM MIDDLE**) Boxplot showing biomarker score stratified by dysplasia grade. (**BOTTOM RIGHT**) Density plot of biomarker scores stratified by dysplasia status.

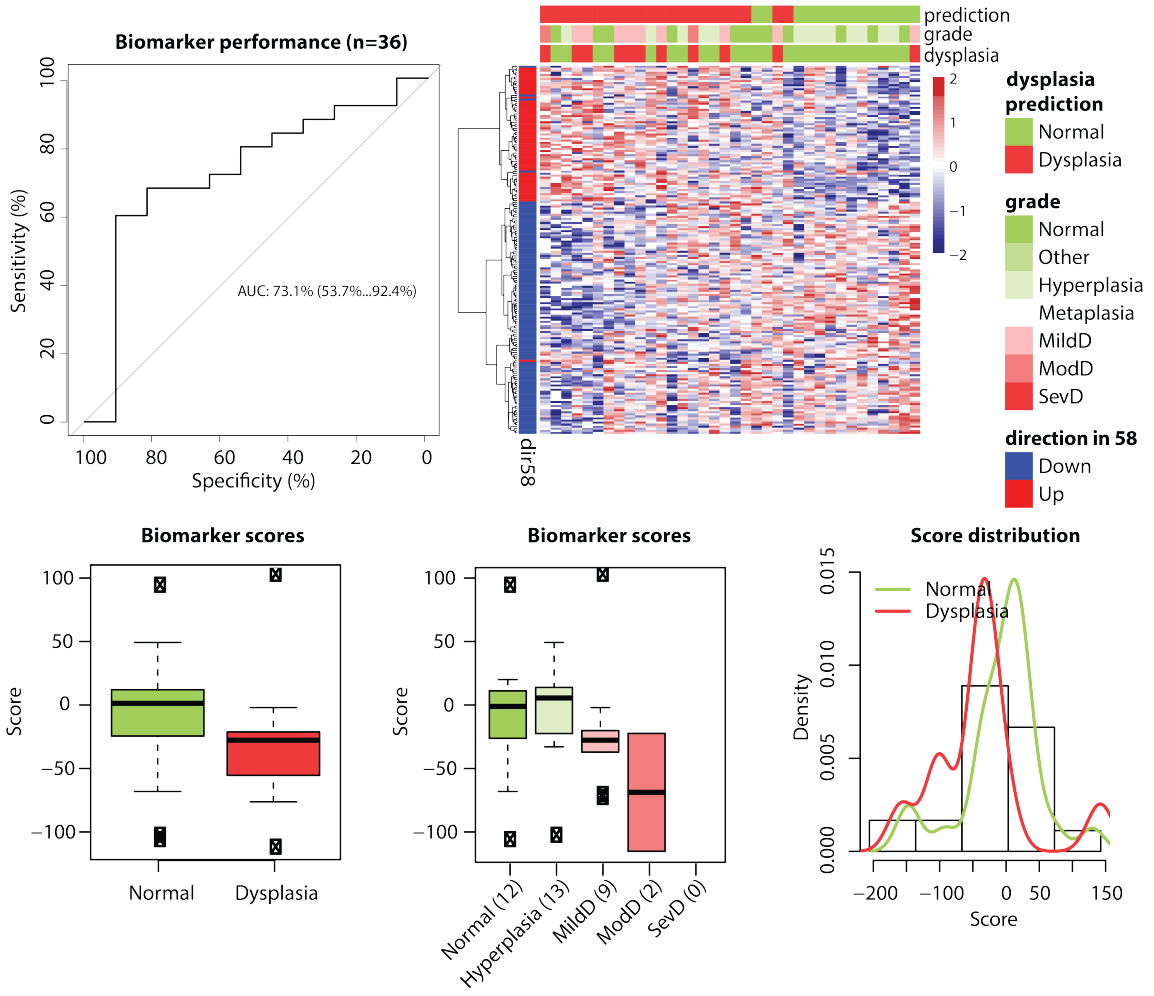


Figure 2.6 Biomarker performance in an overlapping microarray set (n=36).
To evaluate biomarker's performance predicting PML presence, the biomarker was trained on discovery set samples (n=58) and tested in overlapping microarray validation samples (n=36). (**TOP LEFT**) ROC curve summarizing performance in the validation set. (**TOP RIGHT**) Heatmap showing 200 biomarker genes (rows) and 36 validation samples (columns). Horizontal color bars correspond to dysplasia / normal phenotype: top bar shows biomarker prediction, and middle and bottom bars show actual dysplasia grade and status. Vertical color bar corresponds to the directionality of genes in 58 discovery samples. (**BOTTOM LEFT**) Boxplot showing biomarker score stratified by dysplasia status. (**BOTTOM MIDDLE**) Boxplot showing biomarker score stratified by dysplasia grade. (**BOTTOM RIGHT**) Density plot of biomarker scores stratified by dysplasia status.

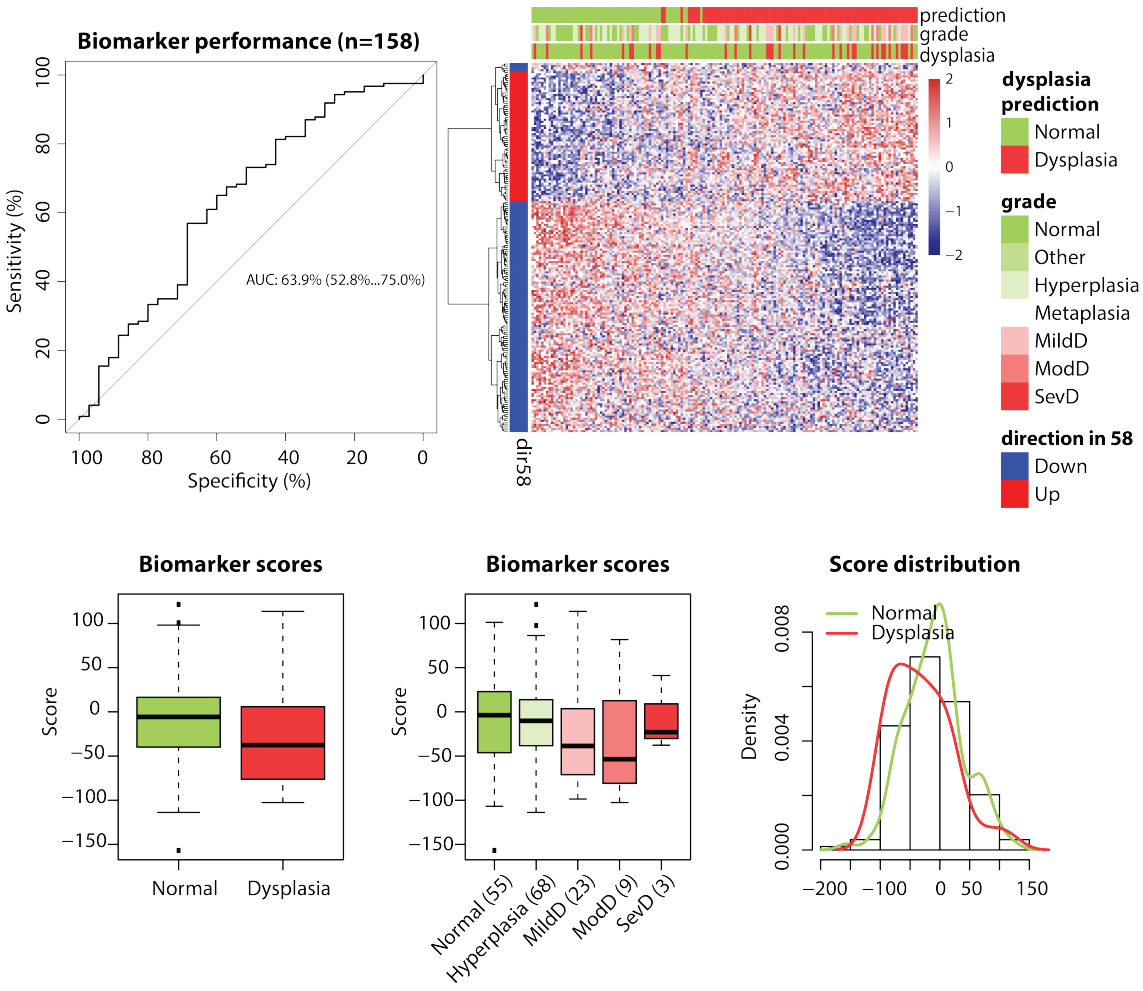


Figure 2.7 Biomarker performance in an independent microarray set (n=158).
To evaluate biomarker's performance predicting PML presence, the biomarker was trained on discovery set samples (n=58) and tested in independent microarray validation samples (n=158). (**TOP LEFT**) ROC curve summarizing performance in the validation set. (**TOP RIGHT**) Heatmap showing 200 biomarker genes (rows) and 158 validation samples (columns). Horizontal color bars correspond to dysplasia / normal phenotype: top bar shows biomarker prediction, and middle and bottom bars show actual dysplasia grade and status. Vertical color bar corresponds to the directionality of genes in 58 discovery samples. (**BOTTOM LEFT**) Boxplot showing biomarker score stratified by dysplasia status. (**BOTTOM MIDDLE**) Boxplot showing biomarker score stratified by dysplasia grade. (**BOTTOM RIGHT**) Density plot of biomarker scores stratified by dysplasia status.

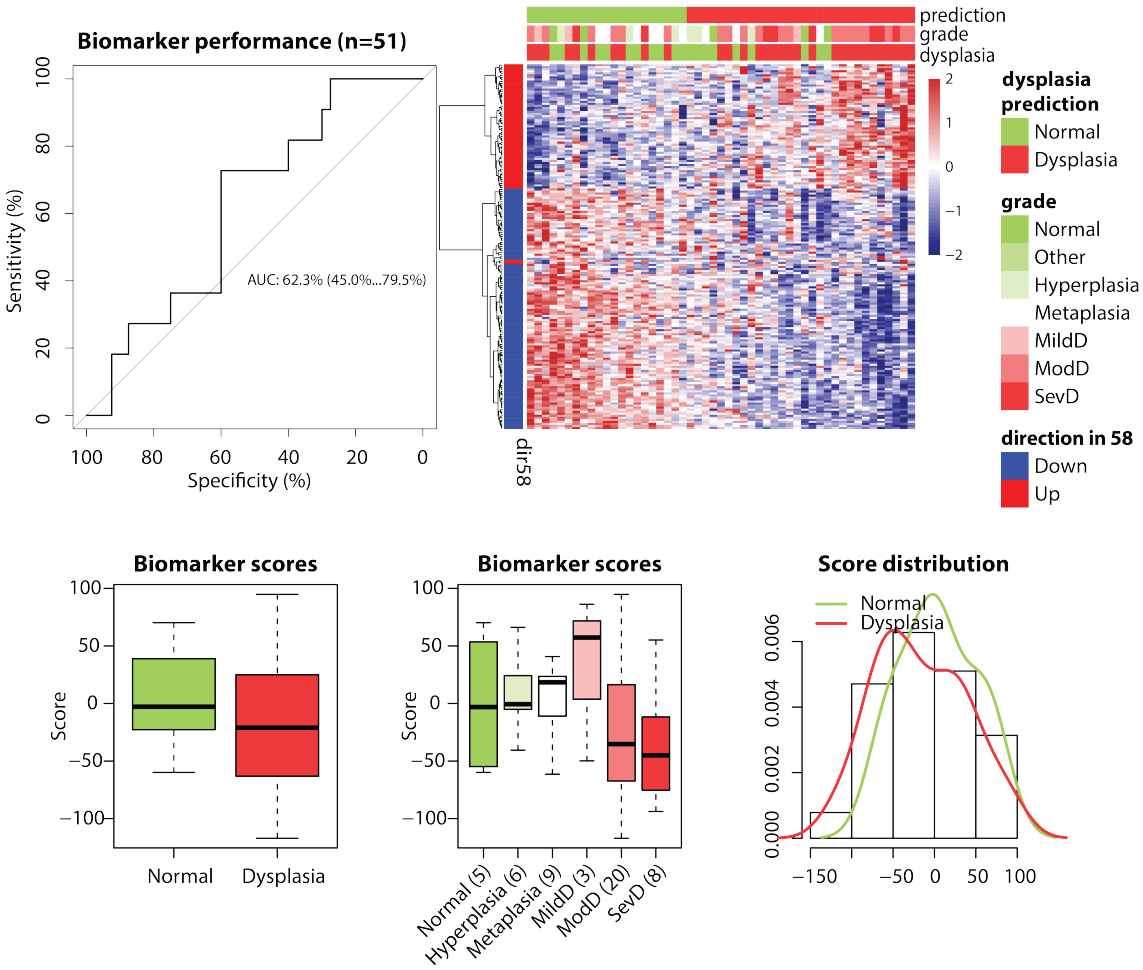


Figure 2.8 Biomarker performance in a longitudinal RNA-Seq dataset (n=51).
To evaluate biomarker's performance predicting PML presence, the biomarker was trained on discovery set samples (n=58) and tested in longitudinally collected RNA-Seq validation samples (n=51). (**TOP LEFT**) ROC curve summarizing performance in the validation set. (**TOP RIGHT**) Heatmap showing 200 biomarker genes (rows) and 51 validation samples (columns). Horizontal color bars correspond to dysplasia / normal phenotype: top bar shows biomarker prediction, and middle and bottom bars show actual dysplasia grade and status. Vertical color bar corresponds to the directionality of genes in 58 discovery samples. (**BOTTOM LEFT**) Boxplot showing biomarker score stratified by dysplasia status. (**BOTTOM MIDDLE**) Boxplot showing biomarker score stratified by dysplasia grade. (**BOTTOM RIGHT**) Density plot of biomarker scores stratified by dysplasia status.

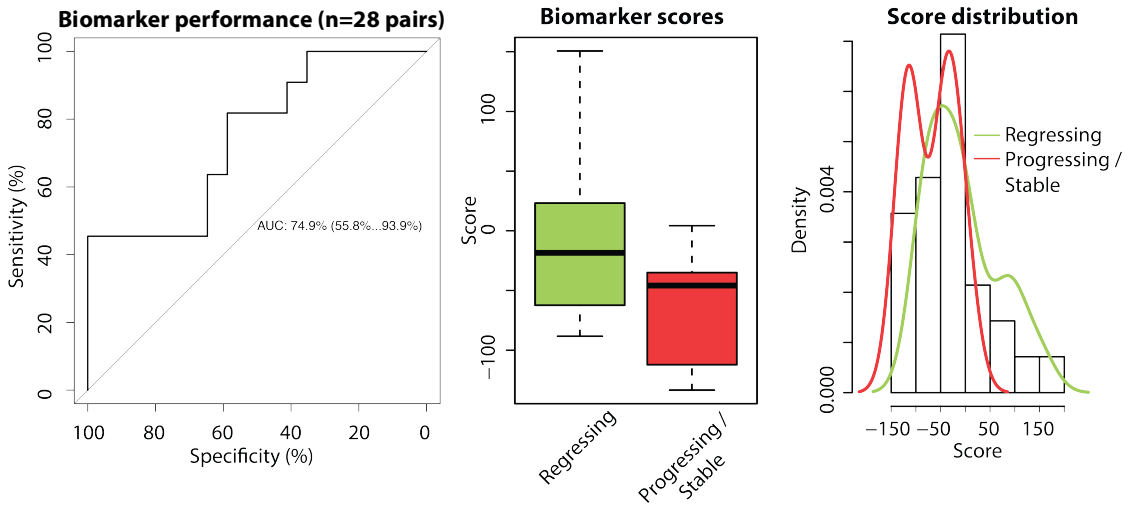


Figure 2.9 Performance of biomarker score differences predicting progressing and stable/regressing PMLs (n=28 pairs).
To evaluate biomarker's performance predicting PML progression, biomarker scores were first calculated independently for each sample (Figure 2.8). Consecutively collected samples were then paired, and the difference between post and pre biomarker scores was calculated. In addition, progression and regression were calculated as the difference in dysplasia grade between post and pre time points. (**LEFT**) ROC curve summarizing performance of biomarker score difference predicting PML progression. (**MIDDLE**) Boxplot showing biomarker score differences stratified by PML progression. (**BOTTOM RIGHT**) Density plot of biomarker score differences stratified by PML progression.

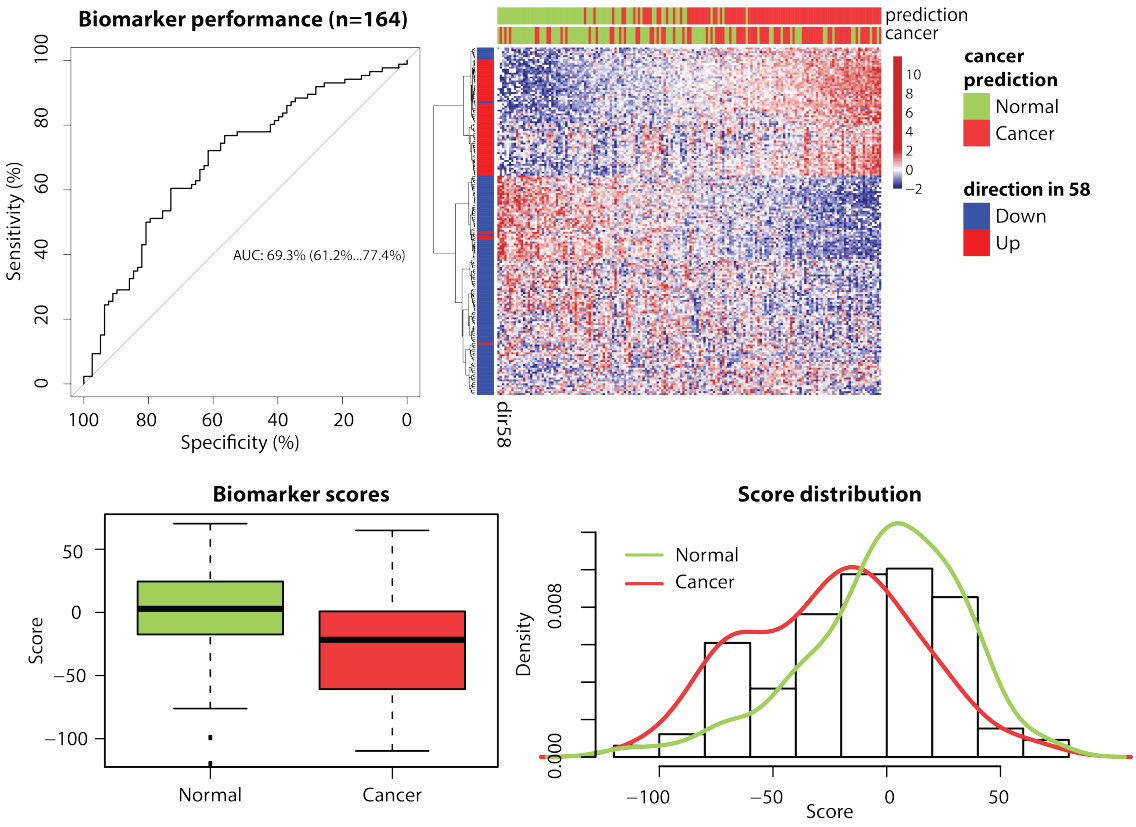


Figure 2.10 Biomarker performance in a microarray lung cancer set 1 (n=164).
To evaluate biomarker's performance predicting lung cancer presence, the biomarker was trained on discovery set samples (n=58) and tested in microarray validation samples (n=164). (**TOP LEFT**) ROC curve summarizing performance in the validation set. (**TOP RIGHT**) Heatmap showing 200 biomarker genes (rows) and 164 validation samples (columns). Horizontal color bars correspond to cancer / normal phenotype: top bar shows biomarker prediction, and bottom bar shows actual lung cancer status. Vertical color bar corresponds to the directionality of genes in 58 discovery samples. (**BOTTOM LEFT**) Boxplot showing biomarker score stratified by cancer status. (**BOTTOM RIGHT**) Density plot of biomarker scores stratified by cancer status.

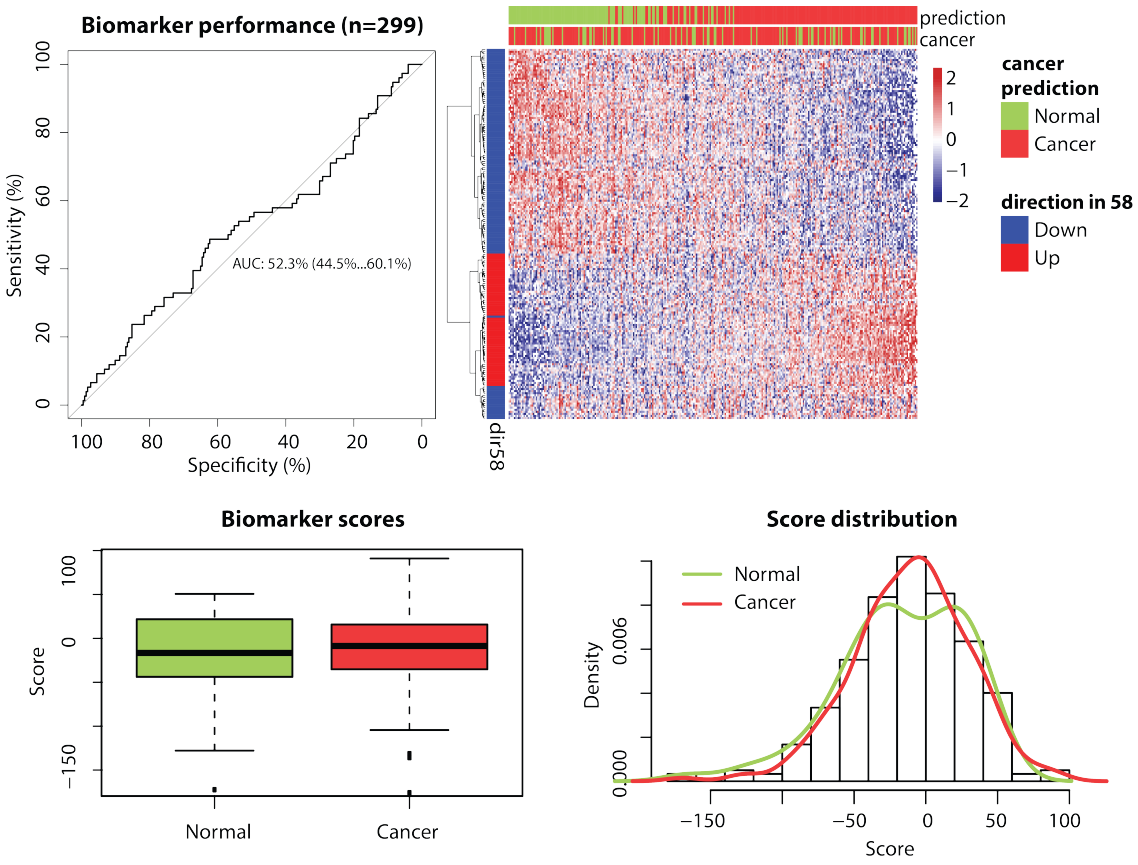


Figure 2.11 Biomarker performance in a microarray lung cancer set 2 (n=299).
To evaluate biomarker's performance predicting lung cancer presence, the biomarker was trained on discovery set samples (n=58) and tested in microarray validation samples (n=299). (**TOP LEFT**) ROC curve summarizing performance in the validation set. (**TOP RIGHT**) Heatmap showing 200 biomarker genes (rows) and 299 validation samples (columns). Horizontal color bars correspond to cancer / normal phenotype: top bar shows biomarker prediction, and bottom bar shows actual lung cancer status. Vertical color bar corresponds to the directionality of genes in 58 discovery samples. (**BOTTOM LEFT**) Boxplot showing biomarker score stratified by cancer status. (**BOTTOM RIGHT**) Density plot of biomarker scores stratified by cancer status.

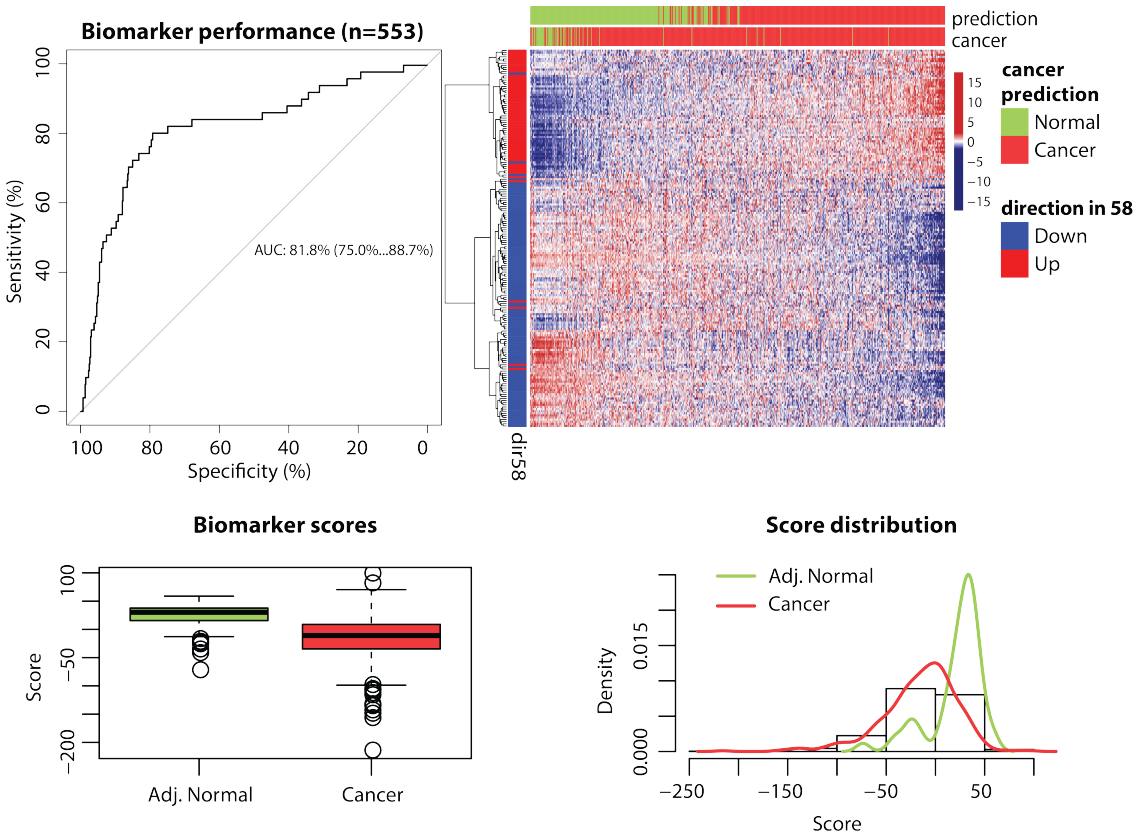


Figure 2.12 Biomarker performance in an RNA-Seq lung SCC tumor biopsy dataset (n=553).
To evaluate biomarker's performance predicting tumor vs. normal samples in subjects with lung cancer, the biomarker was trained on discovery set samples (n=58) and tested in RNA-Seq tumor and adjacent normal biopsy validation samples (n=553). (**TOP LEFT**) ROC curve summarizing performance in the validation set. (**TOP RIGHT**) Heatmap showing 200 biomarker genes (rows) and 553 validation samples (columns). Horizontal color bars correspond to cancer (tumor) / normal phenotype: top bar shows biomarker prediction, and bottom bar shows actual tissue type. Vertical color bar corresponds to the directionality of genes in 58 discovery samples. (**BOTTOM LEFT**) Boxplot showing biomarker score stratified by tissue type. (**BOTTOM RIGHT**) Density plot of biomarker scores stratified by tissue type.

Table 2.12 Functional enrichment of 200 biomarker genes.
Functional enrichment analysis conducted using Enrichr. Shown are categories with significant (q-value < 0.05) enrichment.

database	category	pval	qval
Reactome_2016	Respiratory electron transport, ATP synthesis by chemiosmotic coupling, and heat production by uncoupling proteins._Homo sapiens_R-HSA-163200	1.95E-13	8.94E-11
WikiPathways_2016	Electron Transport Chain_Homo sapiens_WP111	1.51E-12	2.14E-10
KEGG_2016	Huntington's disease_Homo sapiens_hsa05016	8.02E-11	4.02E-09
KEGG_2016	Oxidative phosphorylation_Homo sapiens_hsa00190	5.21E-11	4.02E-09
KEGG_2016	Alzheimer's disease_Homo sapiens_hsa05010	1.12E-10	4.02E-09
KEGG_2016	Parkinson's disease_Homo sapiens_hsa05012	1.26E-10	4.02E-09
GO_Biological_Process_2015	respiratory electron transport chain (GO:0022904)	4.90E-12	5.58E-09
GO_Biological_Process_2015	electron transport chain (GO:0022900)	7.07E-12	5.58E-09
Reactome_2016	The citric acid (TCA) cycle and respiratory electron transport_Homo sapiens_R-HSA-1428517	2.93E-11	6.70E-09
Reactome_2016	Respiratory electron transport_Homo sapiens_R-HSA-611105	1.00E-09	1.53E-07
WikiPathways_2016	Electron Transport Chain_Mus musculus_WP295	2.59E-08	1.83E-06
KEGG_2016	Non-alcoholic fatty liver disease (NAFLD)_Homo sapiens_hsa04932	2.94E-07	7.48E-06
Reactome_2016	Gene Expression_Homo sapiens_R-HSA-74160	1.20E-06	0.000137621
WikiPathways_2016	mRNA Processing_Homo sapiens_WP411	4.54E-06	0.000213493
Transcription_Factor_PPIs	ESR1	1.48E-06	0.000271659
Transcription_Factor_PPIs	PPARGC1A	3.02E-06	0.000276418
GO_Biological_Process_2015	generation of precursor metabolites and energy (GO:0006091)	9.58E-07	0.00050391
WikiPathways_2016	mRNA processing_Mus musculus_WP310	4.04E-05	0.001423104
Transcription_Factor_PPIs	POU5F1	3.07E-05	0.001870846
GO_Biological_Process_2015	hydrogen transport (GO:0006818)	6.63E-06	0.002092956
GO_Biological_Process_2015	proton transport (GO:0015992)	5.86E-06	0.002092956
GO_Biological_Process_2015	ribonucleoprotein complex assembly (GO:0022618)	8.96E-06	0.002355194
Transcription_Factor_PPIs	ESR2	6.73E-05	0.002464967
Transcription_Factor_PPIs	RARA	5.90E-05	0.002464967
KEGG_2016	Cardiac muscle contraction_Homo sapiens_hsa04260	0.000117	0.002482852
GO_Biological_Process_2015	hydrogen ion transmembrane transport (GO:1902600)	1.12E-05	0.00253395
GO_Biological_Process_2015	ribonucleoprotein complex subunit organization (GO:0071826)	1.34E-05	0.002634223
Transcription_Factor_PPIs	HDAC8	0.000109	0.003329301
Transcription_Factor_PPIs	POU4F1	0.000285	0.007445757
Transcription_Factor_PPIs	AR	0.000355	0.008122111
WikiPathways_2016	Oxidative phosphorylation_Homo sapiens_WP623	0.000303	0.008544391
Reactome_2016	Complex I biogenesis_Homo sapiens_R-HSA-6799198	0.000115	0.010571823
Transcription_Factor_PPIs	HTT	0.000652	0.013250879
Transcription_Factor_PPIs	ILF3	0.000724	0.013250879
Transcription_Factor_PPIs	NCOR1	0.000856	0.013263767
Transcription_Factor_PPIs	CEBPA	0.00087	0.013263767
Reactome_2016	TP53 Regulates Metabolic Genes_Homo sapiens_R-HSA-5628897	0.000177	0.013498347
Transcription_Factor_PPIs	YY1	0.001189	0.016731218
KEGG_2016	Metabolic pathways_Homo sapiens_hsa01100	0.0011	0.019965827
Transcription_Factor_PPIs	HDAC2	0.001575	0.020585321
Transcription_Factor_PPIs	ATF2	0.002401	0.027221765
Transcription_Factor_PPIs	POLR2A	0.002529	0.027221765
Transcription_Factor_PPIs	FOXP3	0.002456	0.027221765
Transcription_Factor_PPIs	UBTF	0.003019	0.030694945
Reactome_2016	Formation of ATP by chemiosmotic coupling_Homo sapiens_R-HSA-163210	0.000472	0.030911323
Reactome_2016	Transcriptional activation of mitochondrial biogenesis_Homo sapiens_R-HSA-21	0.00062	0.035512591
Transcription_Factor_PPIs	NR4A2	0.004073	0.03922943
Transcription_Factor_PPIs	AIRE	0.004468	0.040885473
Transcription_Factor_PPIs	SREBF1	0.004708	0.041023352
Transcription_Factor_PPIs	CBX3	0.005328	0.042391957
Transcription_Factor_PPIs	PDX1	0.005217	0.042391957
BioCarta_2016	Pelp1 Modulation of Estrogen Receptor Activity_Homo sapiens_h_pelp1Pathway	0.000937	0.043108539
Transcription_Factor_PPIs	SMARCA4	0.005947	0.043988314
Transcription_Factor_PPIs	MYC	0.00625	0.043988314
Transcription_Factor_PPIs	SP3	0.006076	0.043988314

2.3. Discussion

Lung cancer is the leading cause of cancer death in the US, claiming over 160,000 lives annually. Although CT screening has been shown to be efficacious in detecting potentially cancerous lung nodules [109], it is not commonly accessible and its low specificity leads to overdiagnosis [32,91]which in turn increases patient burden and healthcare cost. The low survivability and high mortality among diagnosed patients underscores the urgent need for early stage diagnostics in lung cancer. One of the features that sometimes accompany lung cancer are bronchial premalignant lesions (PMLs), i.e. histological abnormalities often preceding the development of SCC. Although PML prevalence is fairly low (in a 1999 study, mild, moderate, severe dysplasia and CIS had a prevalence of 44%, 13%, 6%, and 1.6%, respectively [67]), their presence in the airway indicates an increased risk for developing lung cancer [10,127]. Currently, PMLs can be detected using autofluorescence bronchoscopy, which similarly to CT, is not commonly administered, especially in a preventative context. We hypothesized that the airway field of injury developed in cancer-free subjects, can act as a surrogate for the changes observed in the early stages of carcinogenesis by mirroring the presence of PMLs in the airway. To test this hypothesis, we sought to provide a more specific way to detect and monitor PMLs over time, utilizing broadly available and minimally-invasive technology such as white light bronchoscopy combined with gene expression profiling, at an easily accessible site such as the main stem bronchus. Our findings provide novel insights into the earliest molecular events associated with lung carcinogenesis and have the potential

to impact lung cancer prevention by providing additional opportunities to evaluate high-risk smokers at stages of the disease early enough for preventative treatment to be effective, including biomarkers for risk stratification and the means to monitor the efficacy of chemoprevention agents targeted towards the reduction of PML incidence.

In this study, using cytologically normal bronchial airway cells obtained via autofluorescence bronchoscopy from lung cancer-free current and former smokers, we developed a gene expression-based biomarker capable of distinguishing bronchial brushing samples from subjects with and without PMLs. Our biomarker discovery pipeline allowed us to test thousands of models, and select a final signature with the power to detect as well as monitor PMLs over time. Moreover, we identified biological pathways that are dysregulated in the airway field of injury (Table 2.12). Electron transfer chain, formation of ATP, oxidative phosphorylation and metabolism were strongly enriched among genes up-regulated in the airways of subjects with PMLs. Other up-regulated pathways included p53 signaling, and mitochondrial biogenesis [78]. Among upregulated transcription factors, we observed the activating transcription factor 2 (*ATF2*) which is typically overexpressed in NSCLC and is associated with poor prognosis in LC by mediating cell proliferation [145]. Similarly, another implicated transcription factor, the estrogen receptor 1 (*ESR1*), has been shown to have a prognostic value in NSCLC metastasis [7]. These results suggest that the PML-associated airway field harbors shared alterations observed in individuals with lung cancer, which may be interpreted as some of the earliest events in the process of carcinogenesis.

The 200-gene biomarker, measured in the normal-appearing airway epithelium, achieved high performance detecting the presence of dysplastic lesions in a small test set (AUC=0.92). Its high sensitivity suggests that the test could be used in concert with bronchoscopy in cases where lesions cannot be easily visualized due to technological limitations. This could potentially enhance the diagnostic power of bronchoscopy and warrant aggressive follow-up in lung cancer screening programs for high-risk smokers. Additionally, the biomarker might be a fitting candidate for adoption as an intermediate endpoint of efficacy in broader clinical settings including intervention trials. While we were able to verify that the biomarker is capable of discriminating between samples from subjects with and without PMLs in an independent dataset, leveraging the cohort's longitudinal design allowed us to demonstrate an extremely important aspect of this study — the biomarker's capacity to predict progression of the disease. We found that the difference in biomarker scores calculated for pairs of consecutively collected samples, was indicative of regressive or progressive/stable disease as defined by disparities in recorded histology. This result suggests that the PML-associated field of injury changes dynamically over time and that capturing the gene expression longitudinally may allow for further stratification of high-risk smokers based on their associate progression risk.

To further evaluate the biomarker's major utility as a tool to identify at-risk smokers, we examined the behavior of the biomarker genes in overlapping and independent microarray samples. The performance of the biomarker in the overlapping samples was high (AUC=0.73) despite the fact that, historically, biomarkers developed

using RNA-Seq and tested in microarray datasets have, by design, a higher chance of discordance [125]. Although the good performance in these samples could be partially attributed to the fact that the tested samples originated from a subset of subjects used for the development of the biomarker, the biomarker performed not much worse on a subset of independent samples (AUC = 0.63). It is important to note however, that the lower AUC in independent samples may be driven by the low number of severe dysplasias (n=3) with relatively high biomarker scores, which seem to disrupt the overall association of lower biomarker scores with increasing grades of dysplasia (Figure 2.7), thus suggesting that incorporating additional samples represented more equally by each dysplasia grade could help refine the biomarker. Overall, these findings showcase the biomarker's cross-platform applicability, supporting the use of the biomarker in a wider range of settings, especially at medical institutions where gene expression profiling by microarrays is preferred over RNA-Seq for financial or technical reasons.

In addition, we tested the behavior of biomarker genes in the setting of lung cancer, both within bronchial brushings and tumor biopsies. In a microarray brushing dataset 1, as well as TCGA samples, the biomarker predicted the presence of lung cancer exceptionally well (AUC=0.69 in brushings from subjects with and without LC, and AUC=0.87 in lung tumor and adjacent normal biopsies). However, we observed a near random behavior of the biomarker in microarray brushing dataset 2 (AUC=0.52). It is unclear exactly why the biomarker did not perform similarly on the two microarray brushing sets, especially since dataset 2 contained almost twice as many samples as

dataset 1, which eliminates the issue of inadequate sample size to detect differences. However, it is feasible that the different Affymetrix platform used to profile each dataset (HG-U133A for dataset 1 and Human Gene ST 1.0 Array for dataset 2) played a role, and that data preprocessing steps may have to be tailored for the particular array in the future. Another potential explanation may be related to the fact that dysplasia status is unknown for the lung cancer datasets, and has the potential to improve the biomarker's performance in cases were high-grade dysplasia correlates with positive cancer status, and worsen the performance if there is no association. Interestingly, when cancer predictions derived from our biomarker score were compared to those made with an existing clinico-genomic biomarker for lung cancer (PERCEPTA^{TM}) commercialized recently by Veracyte [115]we observed subsets of cancer and normal samples that were predicted correctly only by our biomarker suggesting a potential improvement of overall predictability if the biomarkers are combined that should be explored. Furthermore, we observed that a subset of biological pathways enriched among the biomarker genes were commonly altered in lung cancer, suggesting a similar mechanism of action possibly reflecting the earliest transcriptomic changes in the process of carcinogenesis.

There are a number of limitations to our study. While the size of the discovery set was sufficiently large to warrant statistical power of our classifier, the validation set sample size was relatively small, encouraging the reproduction of our findings in an additional, much larger sequencing dataset in the future. However, we provide extensive evidence for the high degree of validity of the results by testing the biomarker in

additional datasets that differed from the discovery and validation sets by the gene-expression platform used (i.e., RNA-Seq vs. microarray), the condition of interest (i.e., bronchial dysplasia vs. lung cancer), or the tissue sampled (bronchial epithelium vs. SCC tumor). In addition, the histological grade assigned to each brushing was defined as the worst histology observed in multiple biopsies at the same time point within a patient. As such, in theory, it provides a surrogate measure of the overall state of patient's premalignant disease. However, it does not take into consideration several plausible scenarios, including anecdotal cases in which two patients with at least one severe dysplasia lesion have varied composition of the remaining lesions observed (one could have 10 other severe lesions, while the other could have 10 normals). Although intuitively these patients' overall airway health should not be considered to be the same, a phenotype of "severe dysplasia" is likely to be applied to both by currently employed methods, suggesting that further investigation of the effects of the number and severity of PMLs on the airway field of injury should be conducted. Moreover, since the amount of time between procedures is not standardized among study participants, defining progression to best reflect the actual disease evolution, can prove to be a difficult task. The rate at which PMLs change on a histological-grade level (as opposed to condition level, i.e. dysplasia vs. normal) over time remains incompletely understood. Thus, it is unclear if a period of one month between procedures provides enough time to be able to observe a sustained (and real) change in histology; similarly, the merits of defining lesion progression based on two time points 24 months apart might be uncertain. Finally, when

assessing progression, it is important to be aware of the possibility of observing "regression to the mean". In theory, this statistical phenomenon may explain why an extreme case such as severe dysplasia is more likely to regress than remain stable (or progress), and concordantly, why a hyperplastic lesion is more likely to progress. In fact, we observed a significant dependence between histological grade of the baseline lesion and progression in the RPCI data (Pearson's Chi-squared test p-value=0.028). In practice however, the confounding may be due to the fact that performing biopsy on an advanced lesion may have a therapeutic effect by removing most of the abnormal tissue in the process and encouraging wound repair mechanisms and scar formation, which may induce the subsequent (though unsustained) regression of the lesion that may not be reflected by the airway field of injury.

While we were able to demonstrate the utility of the biomarker as a predictor of PML presence in a wide range of datasets, further development and testing in a larger cohort is needed to confirm the biomarker's ability to predict future PML progression. Longitudinal studies would provide the information needed to assess the rate at which the airway field responds to changes in lesion histology and thus help design future studies as well as standardize screening procedures. Moreover, the examination of lesion locations within the respiratory tract could yield insights into the relationship between airway field gene expression and distance to "reporting" lesions. Additionally, alternative methods for classifying brushes based on observed lesion histologies should be explored, as well as the influence of PML number and severity on the overall assessment of premalignant

disease. Finally, important characteristics of many diagnostic procedures, such as negative predictive value (NPV) and positive predictive value (PPV), should be optimized to maximize the efficacy of the biomarker in specific clinical applications, which could potentially change its overall performance.

2.4. Conclusions

In summary, this work marks one of the first attempts at leveraging gene expression harbored by cytologically-normal bronchial epithelial cells to detect and monitor PMLs over time. The high performance of the gene expression-based biomarker in discriminating samples from subjects with and without PMLs provides evidence for the existence of a PML-specific field of injury. In addition, the shared genomic response in the field of injury of subjects with lung cancer and subjects with PMLs who are lung-cancer free, suggests that biological pathways which become activated in diseased subjects (e.g. OXPHOS), may reflect some of the first molecular changes in the process of lung carcinogenesis.

Despite several challenges, we were able to develop a gene expression-based biomarker for the presence of PMLs that suggests great clinical utility. Focusing on lung cancer prevention as opposed to detection by characterizing the transcriptomic events that take place in premalignancy holds great promise and this study represents an important milestone in defining a precision medicine approach to targeting the vulnerable parts in the molecular machinery playing an important role in lung carcinogenesis..

CHAPTER THREE: Identifying miRNA/mRNA Regulatory Interactions Associated with Severity of Lung Cancer Premalignant Lesions

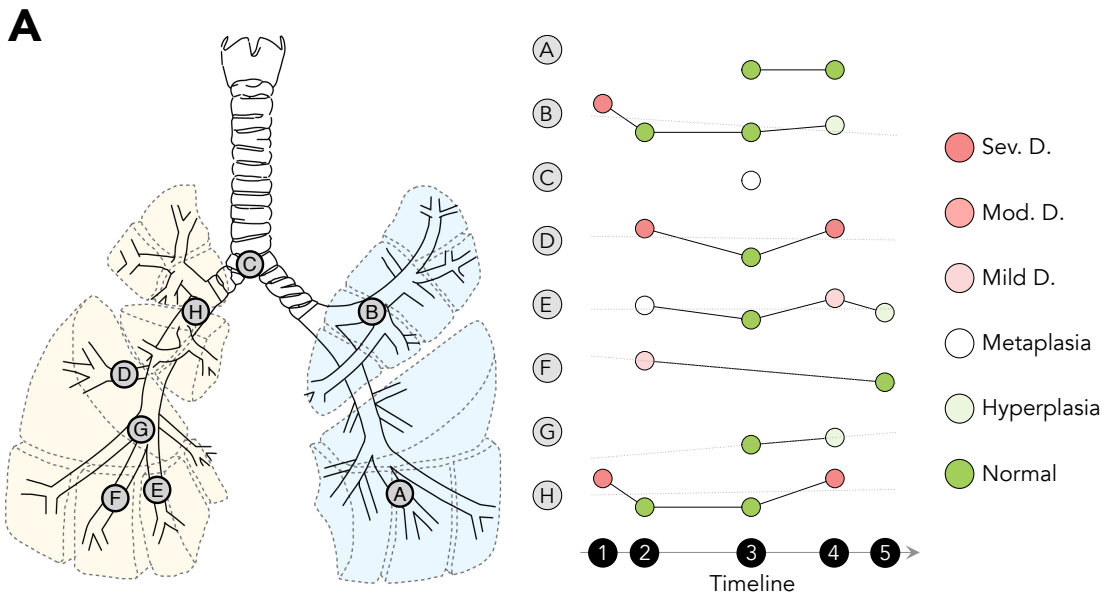
3.1. Background

PMLs are preinvasive histological abnormalities in the central airway, that can be observed and sampled via autofluorescence bronchoscopy, and reproducibly graded by a pathologist [86]. Although PMLs constitute the presumed precursors of SCC of the lung, it's been shown that their multiplicity and severity positively correlate with an increased risk for developing any subtype of LC [10,55]. The natural history of PMLs, which follows a step-wise progression model whereby normal cells proceed through pathological stages from hyperplasia and squamous metaplasia, to mild, moderate and severe dysplasia, to carcinoma in-situ[63,74] does not explain if and when a lesion may progress to invasive SCC. In fact, PMLs are dynamic and their histology might worsen and improve multiple times within a patient [22](Figure 3.1). In this chapter, we hypothesize that monitoring changes in transcriptional and regulatory landscapes of PMLs could allow us to pinpoint the processes involved in the development of premalignant disease and predict PML progression with greater certainty.

The Pre-Cancer Genome Atlas (PCGA) is a promising, multi-consortium initiative, which aims to conduct comprehensive multi-omic longitudinal profiling of PMLs using DNA, mRNA and miRNA sequencing of airway brushes and lesion biopsies [25]. Motivated by the same urgent need to develop alternative premalignancy diagnostics

described in Chapter 2, the PCGA is an invaluable source of sequential bronchoscopies, providing the opportunity to follow the natural history of the PMLs over time *in-vivo* and capture the possibly earliest molecular events that may contribute to the progression of PMLs towards LC.

In deciphering the processes underlying the progression of lesions, we considered the unprecedented role that miRNAs play in gene expression regulation in the context of human disease. miRNAs are short non-coding RNA molecules which target genes for degradation, and it is believed that they regulate nearly all animal processes, including development and pathological processes [45,93]. In cancer, miRNAs can function either as tumor suppressors or oncomiRs, and as such have the potential to influence disease processes by becoming therapeutic targets [93]. In this chapter, we test this theory in the context of premalignant disease by examining gene co-expression networks and their association with miRNAs and the severity and progression of PMLs.



B **Defining progression**
Δ>0 ~ progressing lesion

	H 2	E 4
Long term previous (current grade – worst past grade) [**LT prev**]	regressing	progressing
Short term previous (current grade – preceding grade) [**ST prev**]	regressing	progressing
Short term next (following grade – current grade) [**ST next**]	stable	regressing
Long term next (worst future grade – current grade) [**LT next**]	progressing	regressing
Last biopsy (grade at most recent time point – current gr.) [**last**]	progressing	regressing

Figure 3.1. Example lung map of biopsy locations and corresponding histology grades changing over time.
A real-life example of a patient with a dynamic PML history. (**A LEFT**) Spatial lung map highlighting premalignant lesion locations biopsied at least once for one subject over time. Locations are labeled A-H. (**A RIGHT**) PML histological grade changes across 5 time points. (**B)** Progression definitions and classification of two example lesions (H and E) at selected time points (2 and 4).

3.2. Methods

3.2.1. Sample Collection and Histological Grading

Bronchial lesion biopsies were obtained from current and former smokers enrolled in the High-Risk Lung Cancer-Screening Program at Roswell Park Cancer Institute (RPCI) (Buffalo, NY) between December 2009 and March 2013. Central airways were visualized using autofluorescence bronchoscopy, and abnormally fluorescing areas were sampled and subsequently graded by a pathologist using the following scoring system:

Histology	Dysplasia grade
[0, 22)	Normal
[22, 23)	Hyperplasia
[23, 24)	Metaplasia
[24, 25)	Mild dysplasia
[25, 26)	Moderate dysplasia
[26+)	Severe dysplasia

3.2.2. RNA and miRNA Library Preparation and Sequencing

Total RNA and miRNA was extracted from samples using miRNeasy Mini kit (Qiagen) according to manufacturer's instructions. Sequencing libraries were prepared for total RNA samples using Illumina® TruSeq® RNA Sample Preparation Kit v2 and for miRNA using NEBNext® Multiplex Small RNA Library Prep Set for Illumina (NEBioLabs, Inc.) according to manufacturer’s instructions for each kit. Briefly, the total

RNA was ligated with sequencing adapters, reverse-transcribed, and PCR-amplified to create an individual cDNA library per sample. The libraries were pooled in groups of 6-10 and then PAGE gel-purified to select the adapter-ligated constructs derived from the 22-nt and 30-nt small RNA fragments. The size-selected library pools were then finally sequenced in the corresponding lanes of a Single Flow Cell on the Illumina® HiSeq 2500 to generate more than 10 million single-read 36-bp reads per sample.

3.2.3. Data Generation, Summarization and Quality Control

De-multiplexing and generation of FASTQ files was performed using Illumina® CASAVA. mRNA-Seq reads were aligned to the human genome (hg19) using STAR [34] with default parameters. Alignment and quality metrics were calculated using RSeQC v2.3.3 [140]. Gene expression levels were summarized using RSEM [72]. R software for statistical computing v3.1.1 (http://www.r-project.org) was used to conduct all further analyses. miRNA-Seq reads longer than 15nt were aligned to the hg19 using Bowtie v0.12.7 [69]. miRNA expression levels were quantified using Bedtools v2.9.0 as the number of reads aligning to mature microRNA catalogued in miRBase v18 [42] as previously described in [26]

Sample and Gene Filtering

mRNA and miRNA-Seq data were inspected for the presence of duplicates. Duplicates of two miRNAs with identical expression pattern and ID, but slightly different starting and ending chromosomal positions, were removed from further analysis. Samples

were first filtered based on the availability of mRNA and miRNA sequencing, and only samples sequenced using both protocols were included. mRNA samples were then quality-assessed based on the Transcript Integrity Number (TIN) metric, which measures the degree of *in vitro* RNA degradation [139]. Samples with TIN values beyond 2 standard deviations from the mean TIN value were tagged as potential outliers. Surrogate Variable Analysis was conducted to identify latent sources of variability among miRNA samples using the sva R package. We found that the SV with the strongest contribution was highly correlated with batch. Raw counts were normalized to counts per million (CPM) and log2-transformed using cpm function in edgeR package v3.8.6. Genes unlikely to carry information about traits under investigation due to their very low expression (rowSums [the sum of log2cpm across all samples] < 0) or variability across samples (IQR [Inter-Quartile Range] <= 0) were removed. Library sizes for the QC-filtered raw counts were calculated using the calcNormFactors function in edgeR with the TMM (weighted trimmed mean of M-values) [102] method. Library size-adjusted counts per million (CPM) were log2-transformed. The mean-variance trend of the log2CPM values was then modeled on the observation level using the voom function in limma v3.22.7, taking into account the experimental design including batch and TIN covariates. Principal Component Analysis (PCA) [92] was then conducted on filtered normalized counts, and samples for which either PC1 or PC2 value exceeded 2 standard deviations from the mean of the respective PC were identified as additional potential outliers. Moreover, sample quality was evaluated by clustering samples based on pairwise Pearson's, and any

samples not clustering with the rest were tagged as potential outliers. Samples identified as an outlier by at least two of the evaluated QC measures (TIN, PC1, PC2, and correlation), were removed from downstream analyses. After filtering, raw counts were voom-transformed once again as described above.

3.2.4. Defining Lesion Progression

Lesion progression is typically defined by comparing histology grades assigned to a lesion sampled at two time points. However, the decision regarding which procedure will be considered the baseline and which the follow-up can depend on several factors, including the potential of the PML to "store" information about the past or the future, and the time difference between considered procedures. Taking this into account, we evaluate multiple definitions of progression (Figure 3.1):

LT prev - *Long Term Previous*-based definition of progression - the histology of the lesion under evaluation is compared to the worst histology observed at the same location in the past:

$$\Delta hist = hist_{t_0} - \max_{t<t_0}(hist_t)$$

ST prev - *Short Term Previous* - the histology of the lesion under evaluation is compared to the histology of the lesion observed at an immediately preceding time point:

$$\Delta hist = hist_{t_0} - hist_{t_{-1}}$$

ST next - *Short Term Next* - the histology of the lesion under evaluation is compared to the histology of the lesion observed at an immediately following time point:

$$\Delta hist = hist_{t_{+1}} - hist_{t_0}$$

LT next - *Long Term Next* - the histology of the lesion under evaluation is compared to the worst histology observed at the same location in the future:

$$\Delta hist = \max_{t>t_0}(hist_t) - hist_{t_0}$$

Last - the histology of the lesion under evaluation is compared to the last histology observed at the same location:

$$\Delta hist = hist_{t_{max}} - hist_{t_0}$$

Lesions with a negative histology change were considered regressing, while those for which the change in histology was non-negative were considered stable/progressing.

3.2.5. miRCAT – miRNA Combined Association with Traits

To identify miRNAs and genes associated with traits of interest, we developed miRCAT - miRNA Combined Association with Traits pipeline detailed in following sections. Briefly, miRCAT facilitates the discovery of implicated miRNAs in two ways simultaneously: (1) directly, using linear modeling, and (2) indirectly, by leveraging gene expression (Figure 3.2). Three entities participate in the process: *miRNAs* and their corresponding expression patterns, gene *modules* (groups of coexpressed genes summarized by the 1st principal component), and *traits* including histological lesion grade and the five definitions of lesion progression mentioned above, as well as control traits, such as subtype (molecular subclass derived *de novo* by Dr. Jennifer Beane) and smoking status. In the *direct* step (3.2.11), miRNAs are associated with traits using a linear mixed

effects model, and all miRNAs with a significant FDR are retained for further analysis. In the *indirect* step, miRNAs are first associated with gene modules that contain a significant number of their predicted targets (3.2.10), after which these modules are associated with traits through a linear mixed effects model (3.2.7). Since the function of miRNAs is often determined by the characteristics of genes they target, the direct and indirect steps together allow us to identify not only miRNAs associated with a particular trait, but miRNAs whose predicted targets are also associated and behave like targets.

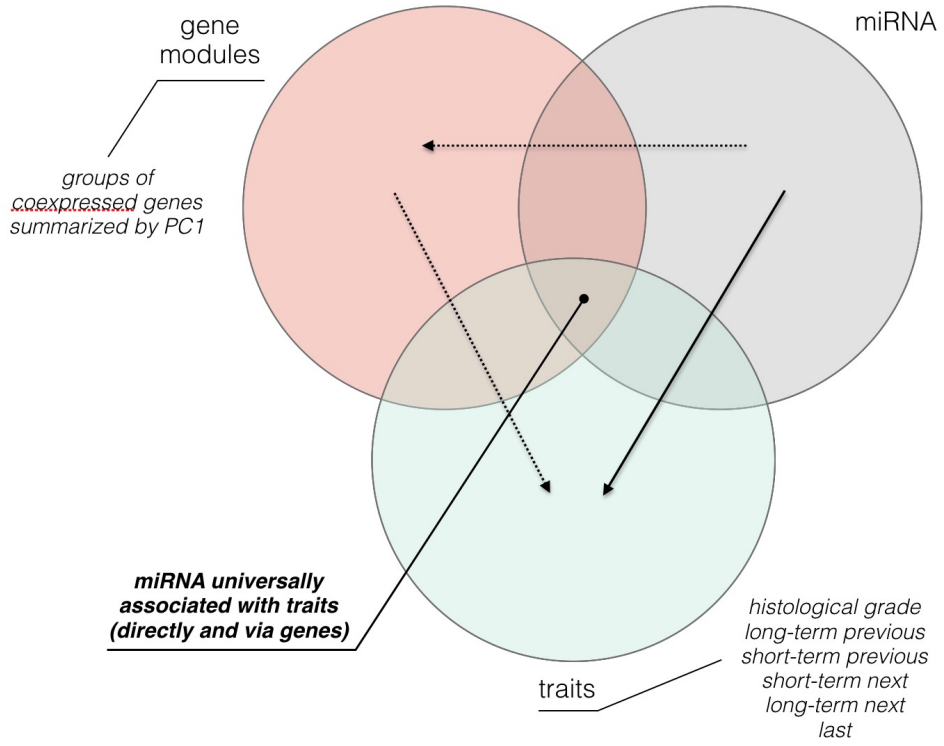


Figure 3.2 miRNA Combined Association with Traits (miRCAT)
Schematic outlining the steps in miRCAT analysis. miRNAs are associated with traits first directly by a linear mixed effects model and then indirectly by trait-association of gene modules containing their targets.

3.2.6. Constructing Gene Coexpression Network

Voom-transformed mRNA reads and the corresponding weights were used to calculate residuals of a linear mixed-effects model using lmer function in lme4 v1.1-8 [13] correcting for fixed quality-related effects such as batch and TIN, and a random patient effect. The Restricted Maximum Likelihood Estimation (REML) feature of the lmer function was turned off in favor of the log-likelihood estimation method. The residuals were then used as input for WGCNA [147].

WGCNA is a systems biology approach which facilitates the identification of subsets of genes that are highly similar to each other. Genes are clustered into disjoint modules based on their expression profile correlation. Module membership and gene connectivity within and between modules can be analyzed to discover novel biological functions, pathways, master regulators or uncovered sources of heterogeneity within samples. Specifically, within the WCGNA framework, a network is defined as a set of nodes (genes) connected by links reflecting the significance of similarity between any pair of gene expression profiles, and genes particularly strongly coexpressed are clustered into disjoint modules. Briefly, an n x m matrix of n genes and m samples is taken as input. First, a simmetrical n x n similarity matrix is filled with values between 0 and 1 corresponding to the absolute values of the Pearson's correlation coefficients calculated between every pair of gene expression profiles:

$$S_{ij} = \left|cor(gene_i, gene_j)\right|$$

The similarity matrix is then transformed into an adjacency matrix. Typically, in an

unweighted network, a threshold is chosen to dichotomize true edges (cor > threshold) and no edges (cor < threshold). In a weighted network however, the measure describing adjacency is continuous, and is obtained by raising the similarity matrix to a soft-thresholding power beta (estimated using the pickSoftThreshold function in the WGCNA v1.46 package), exaggerating the gene-gene relationships (weak relationships become weaker and strong become stronger). The *beta* parameter, or soft threshold, is chosen in a way that ensures the network's scale-free topology (few hubs, many sparsely connected genes). Next, a Topological Overlap Matrix (TOM) is calculated based on adjacency values, which defines the degree of interconnectedness between any pair of genes (two genes are most strongly interconnected if they are each other's neighbors and if they share all of their other neighbors) [144]. Finally, using average linkage hierarchical clustering algorithm, genes are assigned to disjoint clusters (modules). To make module identification and referencing easier, WGCNA assigns unique colors to all modules. Customarily, genes insufficiently co-expressed with any other gene (cor <threshold) are assigned to the grey module.

To ensure the reproducibility of module assignment among genes, WGCNA was additionally ran 50 times in cross validation on randomly chosen 80% of samples using the same *beta* value and seed = 20170217 + iteration number. The `matchLabels` function in WGCNA was used to standardize module color labeling among the reference and all cross-validation runs. Briefly, overlaps between the reference modules and each of the cross-validation assignments were assessed using Fisher's Exact Test, after which

each cross-validation module was labeled with the color of the reference module with which it overlapped most significantly. A consensus network was generated by assigning each gene to its most frequently assigned in cross-validation module. The conservation of modules in the reference network was assessed by module overlaps with the consensus network and any module with FET p-value < 0.05 was considered conserved. It is important to note that the consensus network was used solely to establish preservation of modules in the reference network, and only the unmodified reference network was used in further analysis.

3.2.7. Identifying Genes Associated with PML Grade and Progression

Module-specific gene expression was summarized by the 1st principal components (referred to as module eigengene or ME hereafter) calculated using the expression profiles of module gene members'. MEs facilitate the quantification of co-relationships between modules and clinical variables, as well as assessment of essentiality of each gene as a module member. Due to high correlation of genes within any given module, the ME-based summarization allows to draw accurate conclusions about the association between traits of interest and individual genes.

A linear mixed-effects model was used to assess the association between MEs and traits of interest. Since each module was summarized based on the quality-adjusted residual expression of its members, the fit was adjusted for random patient effect only:

$$Full\ model: ME \sim trait\ +\ 1|patient$$

$$Null\ model: ME \sim 1|patient$$

Module-trait pairs were retained for further analysis, if the FDR-corrected ANOVA p-*value* between full and null models did not exceed the threshold of 0.05.

3.2.8. Defining miRNA Gene Targets

Prior biological knowledge about computationally and experimentally validated miRNA targets was leveraged in this step of the miRCAT algorithm. The publicly available online miRGate database[5] (http://mirgate.bioinfo.cnio.es/miRGate/) was queried using an Application Program Interface (API). The compiled matrix included a row for each miRNA-gene interaction with the corresponding statistical information about the strength of prediction depending on algorithm used (Pictar, miRBase, miRanda, RNAhybrid, TargetScan). For each miRNA, a target set (a list of predicted targets) was then compiled, which included genes identified by at least one of the prediction algorithms.

3.2.9. Defining miRNA "Gene Neighbors"

For each miRNA, a gene "neighborhood" was defined as well. Pairwise Spearman's correlation coefficients were calculated for every miRNA-gene pair and corrected for multiple comparisons using the Bonferroni method. Genes significantly anti-correlated with a miRNA (*rho* < 0, adj.p < 0.05) were included in its "neighborhood".

3.2.10. Identifying miRNA Associated with Gene Modules

An association between miRNA and gene modules was established by assessing module member overlaps with target sets (3.2.8) and "neighborhoods" (3.2.9). Fisher's Exact Test (FET) was used to quantify the significance of the overlaps, and miRNA-module pairs with FDR<0.05 for both tests were retained for further analysis.

3.2.11. Identifying miRNA Associated with PML Grade and Progression

Similar to how association between modules and traits of interest was established in (3.2.7), the "direct" association between miRNAs and traits was evaluated by a linear mixed effects model, correcting for random patient effect only.

$$Full\ model: miRNA \sim trait\ +\ 1|patient$$

$$Null\ model: miRNA \sim 1|patient$$

miRNA-trait pairs were retained for further analysis, if the FDR-corrected ANOVA p-value between the full and null models did not exceed the threshold of 0.05.

The "indirect" association between miRNAs and traits of interest was evaluated by jointly considering the previously established relationships between miRNAs and gene modules (3.2.10), as well as between gene modules and traits (3.2.7). miRNAs were associated with traits of interest "indirectly" if the miRNA and trait were significantly associated with the same module (Figure 3.3).

We focused on miRNAs that associate with certain traits both directly and indirectly, in other words "universally".

3.2.12. Functional Enrichment

The biology of "universally" associated miRNAs was evaluated by a literature search using miRBase database. The functional annotation of the mediating module gene members was conducted using Enrichr [28] and GeneOntology (GO) via GOenrichmentAnalysis function in WGCNA R package.

3.3. Results

3.3.1. Demographic and Clinical Characteristics of Sample Population

A total of 161 premalignant lesions from 27 current and former smokers were biopsied using autofluorescence bronchoscopy at the RPCI. Each subject underwent the procedure between 1 and 7 times (2.7 times on average) and had between 1 and 17 lesions biopsied (6 lesions on average) across all visits. Subjects were on average 59 years old with about 50 pack years, predominantly Caucasian, and equally likely to be female or male. There were no significant differences in these characteristics between normal and dysplastic samples. However, while samples from current smokers were equally distributed between the two groups, there were significantly more samples from former smokers among normal samples (p=0.001) (we did not observe a significant difference in smoking status on a subject level) (Table 3.1).

3.3.2. Gene, miRNA and Sample filtering

57,773 genes and 2,794 miRNAs were included on the sequencing platform. Mixed miRNA IDs of the form MI#_MIMAT# were split into the corresponding Stem-

Loop Sequence ID (MI#) and the Mature miRNA Accession ID (MIMAT#). Among the miRNAs, two (MI0003127_MIMAT0002808 and MI0003127_MIMAT0026606) had duplicates which were subsequently removed. The remaining 2,792 miRNA Accession IDs, mapped to 2,576 miRNA IDs catalogued in miRBase v20.

A total of 288 RNA and 258 miRNA samples were extracted from lesion biopsies and bronchial brushes. Of those, 251matched mRNA-miRNA samples were retained, and subset to 163 biopsies. After filtering out probes with low or non-variable expression, 16,710 genes and 642 miRNAs remained. Samples were then further filtered based on 4 quality measures: TIN, PC1, PC2 and expression correlation. TIN was calculated for mRNA samples ($mean_{TIN}$=77.92, sd_{TIN}=2.34), and 6 samples were tagged as potential outliers. PC1 and PC2 were calculated for mRNA samples ($mean_{PC1}$~0, sd_{PC1}=57.5, $mean_{PC2}$~0, sd_{PC2}=52.6) and miRNA samples ($mean_{PC1}$~0, sd_{PC1}=11.5, $mean_{PC2}$~0, sd_{PC2}=9.4), and 28 samples were tagged as outliers by at least one of the PCs regardless of molecule. No samples were outliers on a dendrogram by expression correlation. Overall, only 2 samples were considered outliers by at least 2 methods and subsequently discarded, leaving 161 samples used in all further analyses (Table 3.1). Finally, the new expression subset was subject to final gene and miRNA variance filtering, which resulted in 16,733 genes and 643 miRNA (580 unique miRBase IDs) considered throughout the following sections.

3.3.3. Genes Associated with PML Grade and Progression

Quality-adjusted residuals of gene expression were clustered using WGCNA, and 14 disjoint modules were identified (excluding the grey module, which included 4 genes and will be excluded from further summaries). Median module size was 906, with the smallest module containing 82 genes and the largest 3690 genes (Table 3.3). Gene expression within modules was collapsed and summarized by the module eigengene (ME), i.e. the first principal component of the module members' gene expression profiles. The top panel of

Figure 3.4 shows a dendrogram showcasing the relationship between MEs, suggesting that some modules (e.g. pink and grey60), and thus the genes they contain, are more similar to each other than to members of other modules. In addition, an annotated heatmap in the bottom panel of the figure shows the strengths of correlations between module MEs.

One-way ANOVA was conducted to identify gene modules significantly associated with traits of interest, including histological grade, histological subtype, and multiple definitions of progression, as well as smoking status (control) by comparing a full linear mixed effects model ($ME \sim trait + 1|patient$) to the null model ($ME \sim 1|patient$). Because modules were derived using expression residuals corrected for batch and TIN, only the patient random effect was added to the models. Ten modules were associated with at least one trait (FDR<0.05), and all traits except progression definitions based on the past (*LT prev*, *ST prev*) and *last* time points were represented by at least one

module, for a total of 25 unique module-trait pairs. Histological grade was associated with the black, brown, cyan, darkturquoise, greenyellow and magenta modules. Future-based short term (*ST next*) progression was associated with the grey60, pink, and red modules. Future-based long term (*LT next*) progression was associated with the grey60 and pink modules (Table 3.4).

Reproducibility of the gene modules was tested by running WGCNA 50 times in cross-validation, using randomly chosen 80% of samples each time, and combining resulting gene clusterings into one consensus module assignment. Gene membership was compared between the original and consensus module assignments by calculating overlaps using FET. All 14 modules achieved a significant overlap, and thus were considered conserved (Figure 3.5).

3.3.4. miRNAs Associated with PML Grade and Progression

First, one-way ANOVA was conducted to identify miRNA "directly" associated with traits by comparing a full linear mixed effects model (*miRNA ~ trait + batch + 1|patient*) to the *null* model (*miRNA ~ batch + 1|patient*). 47 miRNAs were associated "directly" with at least one trait of interest (127 including control traits) (FDR < 0.05), and grade, *LT next* progression, subtype, and smoking were represented by at least one module, for a total of 52 unique miRNA-trait pairs (219 including control traits) (Table 3.5).

Next, to assess the "indirect" associations with trait, miRNAs were first assigned to gene modules. The miRGate database was used to identify computationally predicted miRNA targets. On average, the 580 queried miRBase IDs had 1,440 targets (sd=932) among the 16,733 queried genes. In parallel, for every miRNA, a gene neighborhood was defined as a set of genes significantly anticorrelated with a given miRNA (Spearman's correlation; Bonferroni adjusted p-value < 0.05). On average, miRNAs had 212 neighbors (sd=286). For every module - miRNA pair, Fisher's Exact Test (FET) was performed to assess the overlap between the module and the miRNA's target set and its neighborhood. 481 unique miRNAs were associated with at least one module via targets only (FET FDR < 0.05). 422 miRNAs were associated with at least one module via neighbors only (FET FDR < 0.05). 164 miRNAs were associated with at least one module via targets and neighbors, and thus retained for further analysis (Table 3.6). Finally, gene modules were associated with traits via ANOVA. The results had already been discussed in 3.3.3 above (Table 3.5).

There were 31 miRNAs associated "universally" ("directly" via LME and "indirectly" by leveraging gene expression) with traits of interest (111 including control traits) (Table 3.5, Figure 3.6). Eight unique modules were involved as association mediators. Modules darkturquoise, brown, magenta and black were frequent mediators of miRNA association with grade, although only darkturquoise did not associate with any other trait (Table 3.7), while the remaining three modules shared their association with subtype (Table 3.10). The grey60 and pink modules tended to mediate association with

future-based short-term (*LT next*) progression (Table 3.8). The same two modules, in addition to blue and red, were also shared mediators for smoking and subtype (Table 3.9, Table 3.10).

We also observed that the universal miRNAs clustered into 4 distinct groups (Figure 3.6). Clusters 1 and 4 contained miRNAs associated with grade, and were down- and up-regulated in high-grade dysplasia, respectively. Cluster 2 contained miRNAs upregulated in progressing lesions and low-grade subtypes. Cluster 3 contained miRNAs upregulated in high-grade dysplasia as well as subtype. Surprisingly however, despite the substantial level of redundancy among the miRNA expression profiles within each cluster, we did not observe a significant overlap among the targets of miRNAs that belong to one cluster (Figure 3.11).

3.3.5. Biological Enrichment and Pathway Analysis

Biological enrichment analysis was conducted using the web-based functional annotation tool Enrichr [28] and GeneOntology (GO) (Table 3.11). Module darkturquoise, which only mediated associations with grade, was enriched in genes implicated in cell cycle processes, regulation of *TP53* activity, epithelial cell differentiation, and Hippo signaling pathway. Genes belonging to the progression-associated grey60 and pink modules were involved in cell differentiation and development. Module lightgreen was dominated by genes altered as part of immune response pathways, however it did not significantly associate with any traits (lowest observed FDR=0.49).

Among cluster 1 miRNAs downregulated in high-grade dysplasia, miRNAs 34b and 34c were found to be strongly associated with module darkturquoise which was primarily cell death and apoptosis related. These miRNAs have previously been found to be relevant in these processes in the context of many cancers [85], [49], as they are directly targeted by *TP53* – a transcription factor encoded by the tumor suppressor gene *p53*, which upon DNA damage or stress response may induce cell cycle arrest. Recently, a study showed that overexpression of miR-34b increases cell sensitivity to radiation in NSCLC A549 cell line [9]. Another cluster 1 member, miRNA-4423 has been shown to be downregulated in bronchial epithelium of smokers with lung cancer as a primate-specific regulator of epithelial cell differentiation [94]. In another study, scientists discovered that low expression of miR-92b is indicative of resistance to cisplatin, a common form of chemotherapy often administered in lung cancer [148], which could be the result of targeting genes responsible for DNA repair. Cluster 4 included miRNAs upregulated in high-grade dysplasia. For example, miR-944, whose association was mediated by the ciliogenesis-related magenta module, is an oncomiR. Up-regulated in NSCLC and a target of *ΔNp63*, it up-regulates *p53* expression and is involved in induction of epidermal differentiation [64]. Recently, a study has demonstrated its potential as an early lung cancer biomarker [98].

Cluster 2 associated with progression and subtype, contained miRNAs mediated by the grey60, pink and blue modules. These modules contained genes implicated in MAPK pathway (responsible for cell differentiation and proliferation in *Ras*-mutated lung cancer), Wnt signaling (related to processes promoting tumor progression [97]),

epithelial cell proliferation, and lung development. Among cluster 2 miRNAs, miR-423 was especially interesting because of existing evidence for its capacity to promote tumor progression[43]. In addition, let-7, a well known tumor suppressor, which is frequently lost in NSCLC, has been shown to reverse tumor progression in mice [131].

Finally, cluster 3 was represented by miRNAs upregulated in high-grade dysplasia and subtype, as well as the magenta and brown modules also mediating associations among cluster 2 genes. A member of cluster 3, miR-136 was shown to exhibit antiviral behavior against Influenza exposed to A549 human lung epithelial cell line [149] and was routinely overexpressed in lung cancer, suggesting a role in inhibition of tumor suppressors. In breast cancer, miR-655 was shown to halt the cell transition from epithelial to mesenchymal phenotype, typical of metastatis.

Table 3.1 Demographic characteristics of the RNA-Seq PML biopsy dataset stratified by dysplasia status.

Data are means (SD) for continuous variables and counts for dichotomous variables. Top table summarizes data by samples, and bottom table summarizes data by subjects. P-values are for the comparison between the Dysplasia and Normal groups, using two sample t-tests for continuous variables or Fisher's exact test for categorical variables.

Factor	Overall	Dysplasia	Normal	p-value
No. Samples	161	66	95	
Age	57.50 (6.22)	57.66 (6.62)	57.4 (5.97)	0.783
Sex				0.87
female	101	42	59	
male	60	24	36	
Race				1
Caucasian	156	64	92	
Other	5	2	3	
Lung Cancer	0	0	0	
Smoking				0.001
current	86	45	41	
former	73	19	54	
NA	2	2	0	
Pack Years	48.91 (19.5)	50.14 (16.56)	48.06 (21.36)	0.509
Histological grade				*< 0.001*
Normal	29		29	
Hyperplasia	26		26	
Metaplasia	40		40	
Mild Dysplasia	19	19		
Moderate Dysplasia	35	35		
Severe Dysplasia	12	12		

Factor	Overall
No. Subjects	27
No. of procedures (mean [range])	2.7 (1-7 visits)
No. of lesions biopsied in total (mean [range])	6 (1-17 lesions)
Days between visits (mean [range])	350 (98-742 days)
Age	59.21 (7.73)
Sex	
female	14
male	13
Race	
Caucasian	25
African American	2
Lung Cancer	0
Smoking	
current	11
former	16
Pack Years	50.44 (22.9)

Table 3.2 Sample quality control results.
Samples were quality-evaluated based on transcript integrity number (TIN), the 1st and 2nd principal components (PC1 and PC2) and correlation with other samples (cor). Samples were excluded if they were considered outliers by at least two quality metrics.

Sample ID	TIN	PC1	PC2	cor	excluded
PCGA-01-0001-070-19300-00587BX	TRUE	FALSE	FALSE	FALSE	FALSE
PCGA-01-0001-074-19762-00901BX	TRUE	TRUE	FALSE	FALSE	TRUE
PCGA-01-0001-077-19762-00899BX	TRUE	TRUE	FALSE	FALSE	TRUE
PCGA-01-0007-050-18131-00528BX	TRUE	FALSE	FALSE	FALSE	FALSE
PCGA-01-0014-089-20660-00744BX	TRUE	FALSE	FALSE	FALSE	FALSE
PCGA-01-0021-010-19826-00723BX	TRUE	FALSE	FALSE	FALSE	FALSE
PCGA-01-0001-037-20126-01120BX	FALSE	TRUE	FALSE	FALSE	FALSE
PCGA-01-0002-050-23843-00279BX	FALSE	TRUE	FALSE	FALSE	FALSE
PCGA-01-0002-050-24062-00443BX	FALSE	TRUE	FALSE	FALSE	FALSE
PCGA-01-0004-050-27317-00606BX	FALSE	TRUE	FALSE	FALSE	FALSE
PCGA-01-0007-050-18033-00456BX	FALSE	TRUE	FALSE	FALSE	FALSE
PCGA-01-0008-050-22263-01196BX	FALSE	TRUE	FALSE	FALSE	FALSE
PCGA-01-0023-010-21429-00830BX	FALSE	TRUE	FALSE	FALSE	FALSE
PCGA-01-0031-050-24285-00856BX	FALSE	TRUE	FALSE	FALSE	FALSE
PCGA-01-0008-050-21682-00841BX	FALSE	TRUE	FALSE	FALSE	FALSE
PCGA-01-0015-011-17638-01327BX	FALSE	TRUE	FALSE	FALSE	FALSE
PCGA-01-0023-048-21429-00829BX	FALSE	TRUE	FALSE	FALSE	FALSE
PCGA-01-0026-089-22473-00859BX	FALSE	TRUE	FALSE	FALSE	FALSE
PCGA-01-0031-021-24726-01062BX	FALSE	TRUE	FALSE	FALSE	FALSE
PCGA-01-0001-074-20126-01116BX	FALSE	FALSE	TRUE	FALSE	FALSE
PCGA-01-0001-074-20343-01315BX	FALSE	FALSE	TRUE	FALSE	FALSE
PCGA-01-0022-089-21996-01170BX	FALSE	FALSE	TRUE	FALSE	FALSE
PCGA-01-0031-050-24061-00666BX	FALSE	FALSE	TRUE	FALSE	FALSE
PCGA-01-0001-050-19762-00903BX	FALSE	FALSE	TRUE	FALSE	FALSE
PCGA-01-0005-050-17489-00728BX	FALSE	FALSE	TRUE	FALSE	FALSE
PCGA-01-0021-050-20449-01092BX	FALSE	FALSE	TRUE	FALSE	FALSE
PCGA-01-0021-070-20449-01089BX	FALSE	FALSE	TRUE	FALSE	FALSE
PCGA-01-0022-093-21996-01171BX	FALSE	FALSE	TRUE	FALSE	FALSE
PCGA-01-0025-074-21689-01105BX	FALSE	FALSE	TRUE	FALSE	FALSE
PCGA-01-0025-094-21689-01107BX	FALSE	FALSE	TRUE	FALSE	FALSE
PCGA-01-0029-048-22678-00815BX	FALSE	FALSE	TRUE	FALSE	FALSE
PCGA-01-0029-093-22678-00813BX	FALSE	FALSE	TRUE	FALSE	FALSE

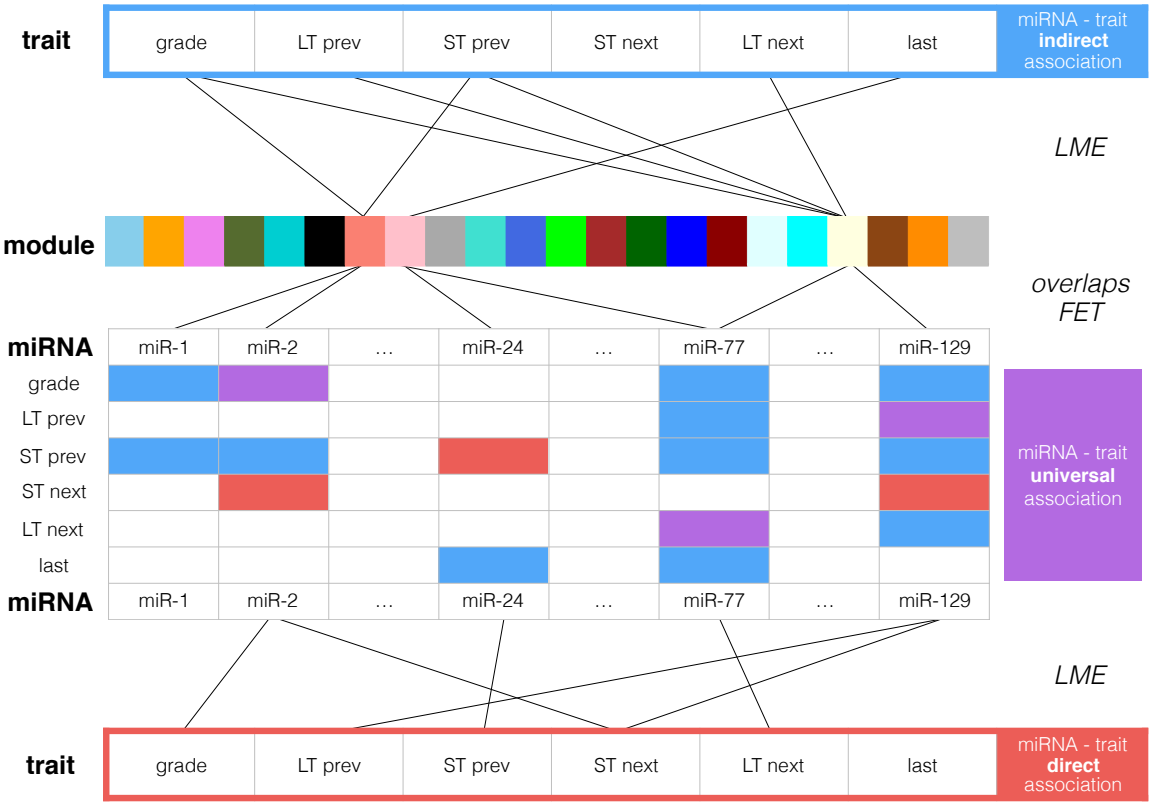


Figure 3.3 Conceptual representation of miRCAT results.
The blue bar on top represents an indirect association with traits. The colorful bar underneath it represent the WGCNA modules. The red bar on the bottom represents direct association with traits. The table visualizes direct (red), indirect (blue) and universal (purple) associations between miRNAs (columns) and traits (rows), Lines connecting traits to modules, modules to miRNAs and miRNAs to traits represent significant associations. miRNAs are considered to be associated universally if they associate both directly and indirectly. miRNA names and module colors are arbitrarily chosen to aid in visualization of the concept, and do not correspond to results.

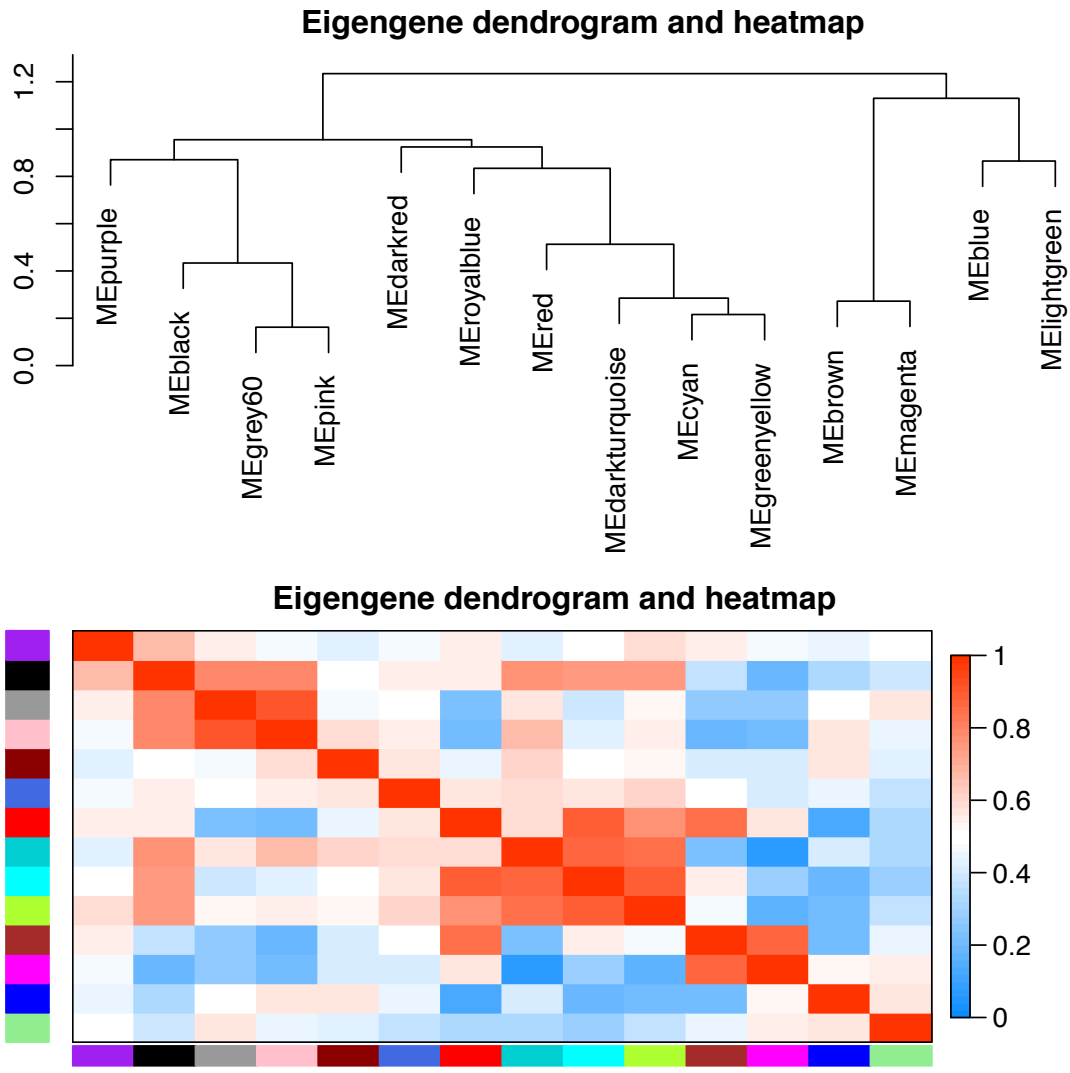


Figure 3.4 Summarized module eigengene relationships.
(**TOP**) Module eigenegene (ME) dendrogram. ME distances quantified with Spearman's correlation. (**BOTTOM**) A heatmap showing the absolute value of correlation between MEs.

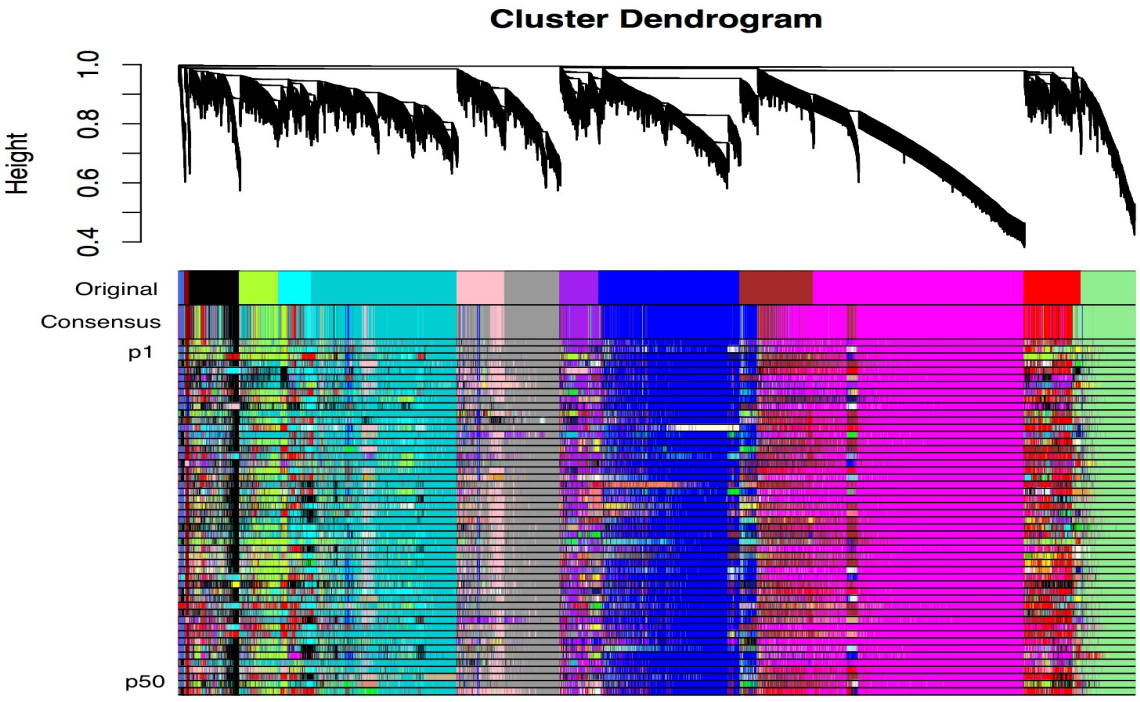


Figure 3.5 Consensus WGCNA.
To evaluate robustness of biopsy-derived mRNA coexpression modules (Original), WGCNA was run 50 times on randomly selected 80% of the samples (p1, p2, …, p50). In each iteration, resulting modules were compared to the original assignment and recolored to reflect strongest overlap with one original module. Consensus modules were generated by assigning each gene to the module most frequently assigned in the 50 iterations.

Table 3.3 WGCNA gene coexpression module sizes

magenta	arkturquois	blue	brown	red	grey60	lightgreen	black
3691	2550	2471	1287	993	970	932	880

pink	purple	greenyellow	cyan	royalblue	darkred	grey
832	680	679	579	103	82	4

Table 3.4 Gene modules significantly associated with traits.
Modules associated with at least one trait of interest and the corresponding significance levels. All listed associations are significant (ANOVA FDR<0.05). The black, brown, cyan, darkturquoise, greenyellow and magenta modules were associated with grade. The grey60 and pink modules were associated with future-based progression. Additionally, the red module specifically associated with short-term future-based progression.

module	trait	value.p	value.fdr
black	grade	1.20E-03	6.01E-03
black	subtype	2.59E-06	7.76E-06
blue	smoking	3.49E-03	1.05E-02
blue	subtype	6.67E-03	1.25E-02
brown	grade	1.95E-03	7.30E-03
brown	subtype	9.07E-07	3.40E-06
cyan	grade	8.81E-03	2.20E-02
cyan	smoking	8.63E-04	3.24E-03
cyan	subtype	7.60E-04	1.63E-03
darkturquoise	grade	1.17E-05	8.79E-05
greenyellow	grade	5.75E-03	1.73E-02
greenyellow	smoking	7.99E-03	2.00E-02
grey60	ST next	7.87E-03	4.46E-02
grey60	LT next	5.14E-04	7.71E-03
grey60	smoking	3.68E-04	2.28E-03
grey60	subtype	4.94E-16	3.70E-15
magenta	grade	1.47E-06	2.21E-05
magenta	subtype	7.32E-04	1.63E-03
pink	ST next	4.66E-03	4.46E-02
pink	LT next	3.74E-03	2.80E-02
pink	smoking	4.56E-04	2.28E-03
pink	subtype	5.70E-18	8.54E-17
red	ST next	8.93E-03	4.46E-02
red	smoking	2.61E-04	2.28E-03
red	subtype	2.54E-09	1.27E-08

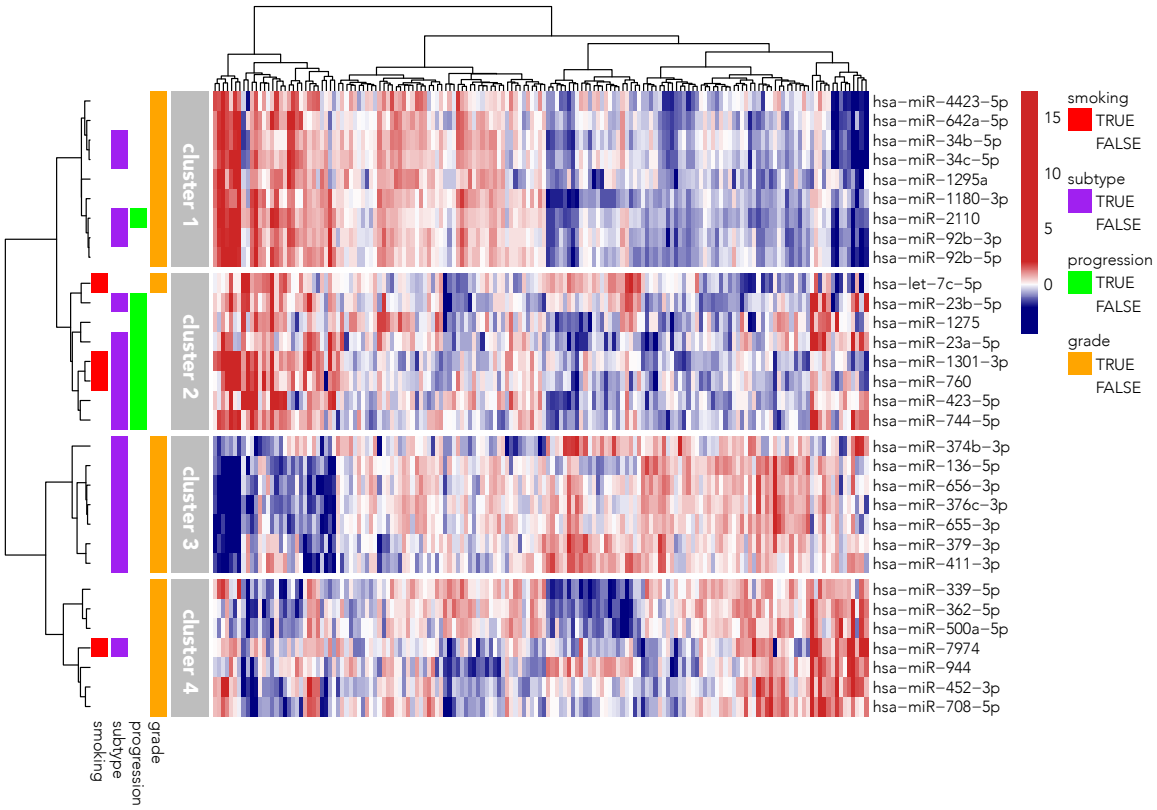


Figure 3.6 Expression of miRNAs universally associated with grade or progression
The unsupervised heatmap clusters miRNAs using Ward.D2 clustering into four groups: clusters 1 and 4 which tend to associate with grade, cluster 2 which associates with progression and subtype, and cluster 3 which associates with grade and subtype.

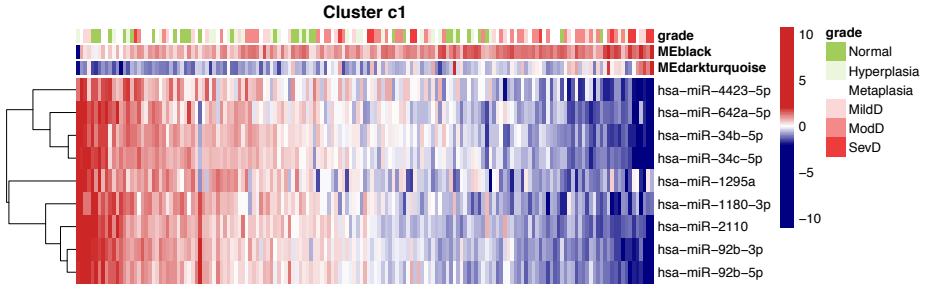


Figure 3.7 Grade-associated miRNAs from cluster 1
Cluster 1 containing 9 miRNAs with an association with grade mediated by the black and darkturquoise modules.

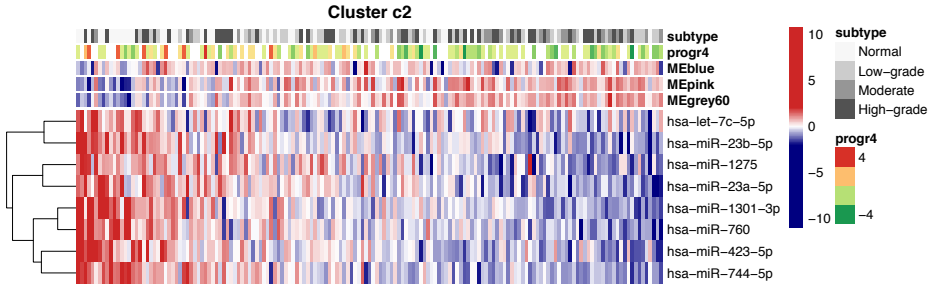


Figure 3.8 Progression and subtype-associated miRNAs from cluster 2
Cluster 2 containing 8 miRNAs with an association with progression and subtype mediated by the blue, pink and grey60 modules.

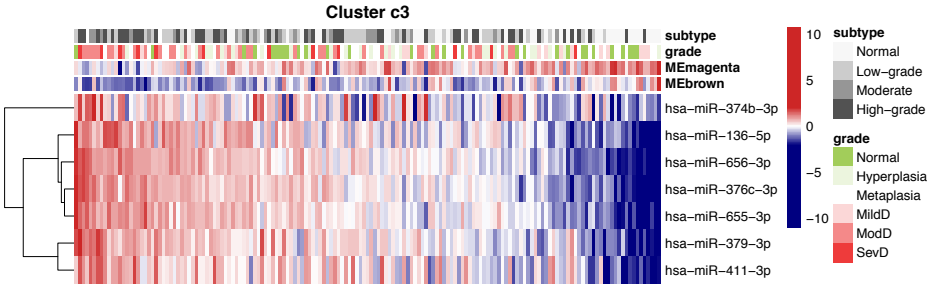


Figure 3.9 Grade and subtype-associated miRNAs from cluster 3
Cluster 3 containing 7 miRNAs with an association with grade and subtype mediated by the magent and brown modules.

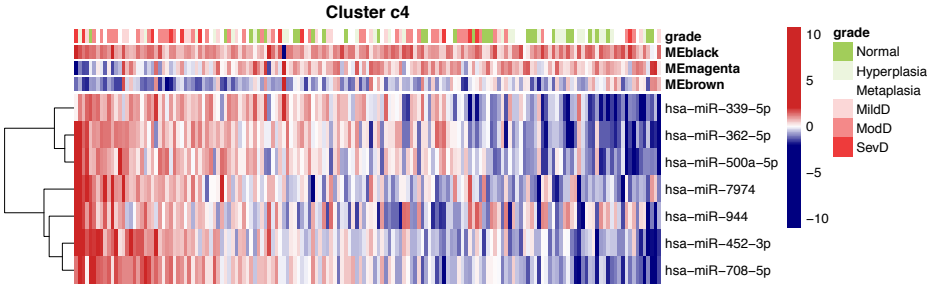


Figure 3.10 Grade-associated miRNAs from cluster 4
Cluster 4 containing 7 miRNAs with an association with grade mediated by the black, magenta and brown modules.

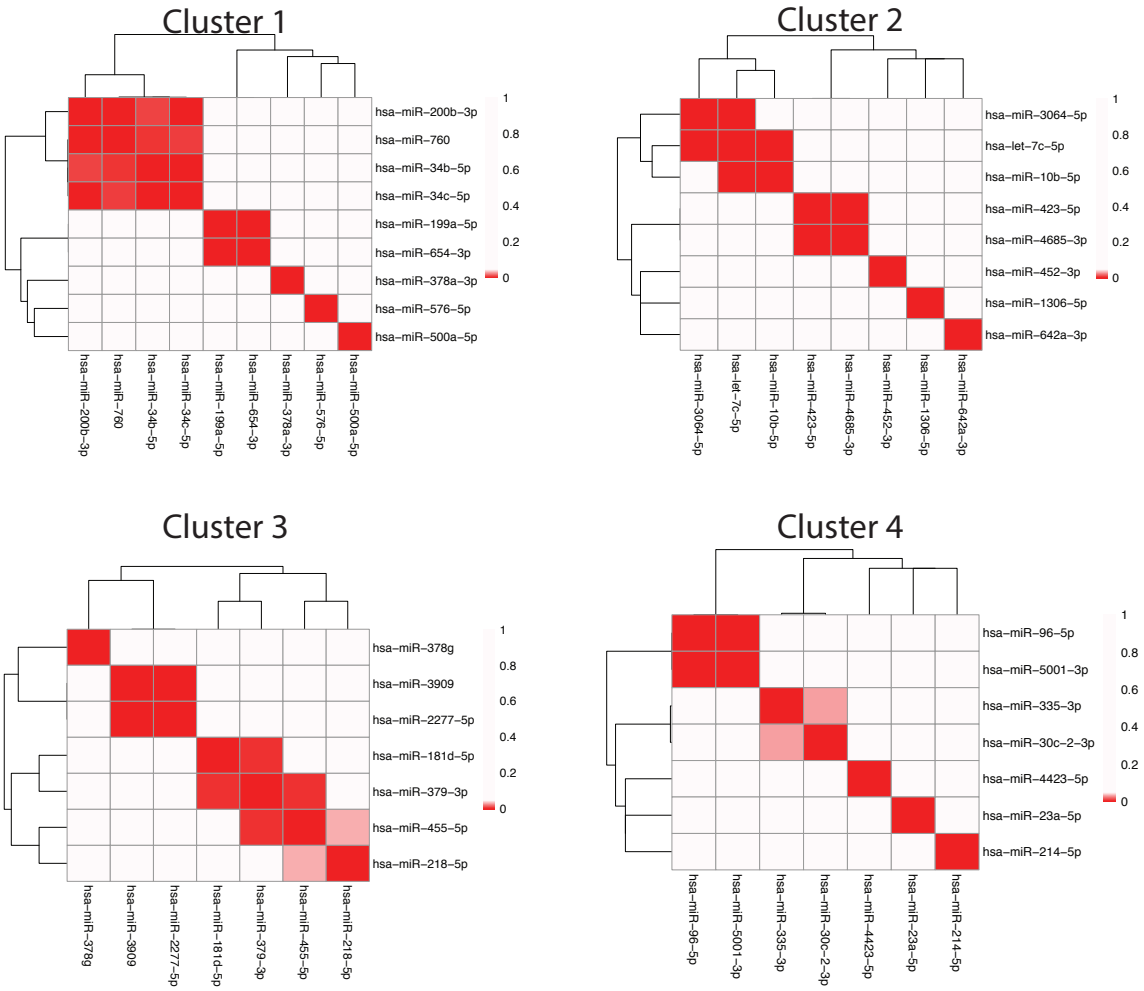


Figure 3.11 Target overlaps in clusters containing coexpressed miRNAs.
Despite high coexpression of miRNAs within these clusters, the miRNAs did not share a significant number of their targets. Color bar corresponds to significance level of Fisher's Exact Test.

Table 3.5 Summary of miRNAs significantly associated with traits of interest.
Number of miRNAs that significantly associate with traits of interest via one-way ANOVA correcting for batch and random patient effect (FDR<0.05). The universal group includes miRNAs associated with a given trait both directly (via ANOVA) or indirectly (via association with gene modules).

	trait	grade	progression					subtype	smoking
			LT prev	ST prev	ST next	LT next	last		
direct	**miRNA**	36	0	0	0	16	0	109	58
indirect	**miRNA**	81	0	0	47	45	0	130	87
universal	**miRNA**	24	0	0	0	8	0	97	34

Table 3.6 Summary of miRNAs associated with gene modules via target sets and "neighborhoods".
Number of miRNAs whose target sets and "neighbors" significantly overlap with modules (Fisher's Exact Test Bonferroni-adjusted p-value < 0.05)

	magenta	darkturquoise	lightgreen	grey60	purple	red	greenyellow
targets	245	299	81	100	22	65	173
neighbors	113	36	89	113	20	58	27
both	42	20	17	25	0	2	2

	blue	brown	black	cyan	pink	royalblue	darkred
targets	263	191	85	0	125	0	0
neighbors	112	65	85	22	107	2	0
both	48	17	10	0	26	0	0

Table 3.7 miRNAs and the mediating modules universally associated with grade.

miRbaseID	trait	module	miRbaseID	trait	module
hsa-let-7c-5p	grade	darkturquoise	hsa-miR-411-3p	grade	magenta
hsa-miR-1180-3p	grade	darkturquoise	hsa-miR-4423-5p	grade	darkturquoise
hsa-miR-1295a	grade	darkturquoise	hsa-miR-452-3p	grade	magenta
hsa-miR-136-5p	grade	brown	hsa-miR-500a-5p	grade	magenta
hsa-miR-136-5p	grade	magenta	hsa-miR-500a-5p	grade	black
hsa-miR-2110	grade	darkturquoise	hsa-miR-642a-5p	grade	darkturquoise
hsa-miR-339-5p	grade	brown	hsa-miR-655-3p	grade	magenta
hsa-miR-34b-5p	grade	darkturquoise	hsa-miR-656-3p	grade	magenta
hsa-miR-34b-5p	grade	black	hsa-miR-708-5p	grade	magenta
hsa-miR-34c-5p	grade	darkturquoise	hsa-miR-7974	grade	brown
hsa-miR-362-5p	grade	magenta	hsa-miR-7974	grade	magenta
hsa-miR-374b-3p	grade	magenta	hsa-miR-92b-3p	grade	black
hsa-miR-376c-3p	grade	magenta	hsa-miR-92b-5p	grade	darkturquoise
hsa-miR-379-3p	grade	magenta	hsa-miR-944	grade	magenta

Table 3.8 miRNAs and the mediating modules universally associated with progression.
LT next is long-term next progression. Some miRNAs are listed multiple times due to the fact that their association with progression was mediated by multiple modules.

miRbaseID	trait	module
hsa-miR-1275	LT next	grey60
hsa-miR-1301-3p	LT next	pink
hsa-miR-1301-3p	LT next	grey60
hsa-miR-2110	LT next	grey60
hsa-miR-2110	LT next	pink
hsa-miR-23a-5p	LT next	grey60
hsa-miR-23b-5p	LT next	pink
hsa-miR-423-5p	LT next	grey60
hsa-miR-423-5p	LT next	pink
hsa-miR-744-5p	LT next	grey60
hsa-miR-760	LT next	grey60

Table 3.9 miRNAs and mediating modules universally associated with smoking
Some miRNAs are listed multiple times due to the fact that their association with smoking was mediated by multiple modules.

miRbaseID	trait	module	miRbaseID	trait	module
hsa-let-7c-5p	smoking	blue	hsa-miR-3065-5p	smoking	pink
hsa-miR-10a-5p	smoking	blue	hsa-miR-30c-2-3p	smoking	blue
hsa-miR-10b-5p	smoking	blue	hsa-miR-330-5p	smoking	grey60
hsa-miR-125a-5p	smoking	blue	hsa-miR-378a-3p	smoking	blue
hsa-miR-125b-1-3p	smoking	blue	hsa-miR-378c	smoking	grey60
hsa-miR-126-5p	smoking	red	hsa-miR-378g	smoking	blue
hsa-miR-127-3p	smoking	blue	hsa-miR-378i	smoking	blue
hsa-miR-1301-3p	smoking	grey60	hsa-miR-422a	smoking	blue
hsa-miR-1301-3p	smoking	pink	hsa-miR-5001-3p	smoking	pink
hsa-miR-147b	smoking	grey60	hsa-miR-548e-3p	smoking	pink
hsa-miR-182-5p	smoking	grey60	hsa-miR-625-5p	smoking	pink
hsa-miR-183-5p	smoking	blue	hsa-miR-671-5p	smoking	grey60
hsa-miR-1976	smoking	grey60	hsa-miR-675-5p	smoking	grey60
hsa-miR-1976	smoking	pink	hsa-miR-760	smoking	grey60
hsa-miR-199a-5p	smoking	blue	hsa-miR-7974	smoking	pink
hsa-miR-200c-3p	smoking	pink	hsa-miR-93-3p	smoking	pink
hsa-miR-2277-5p	smoking	blue	hsa-miR-941	smoking	blue
hsa-miR-3064-5p	smoking	blue	hsa-miR-96-5p	smoking	grey60

Table 3.10 miRNAs and mediating modules universally associated with subtype
Some miRNAs are listed multiple times due to the fact that their association with subtype was mediated by multiple modules.

miRbaseID	trait	module
hsa-let-7f-5p	subtype	blue
hsa-miR-106b-3p	subtype	pink
hsa-miR-106b-5p	subtype	black
hsa-miR-10a-5p	subtype	brown
hsa-miR-10a-5p	subtype	blue
hsa-miR-10b-5p	subtype	brown
hsa-miR-10b-5p	subtype	blue
hsa-miR-125a-5p	subtype	blue
hsa-miR-125a-5p	subtype	brown
hsa-miR-125b-1-3p	subtype	blue
hsa-miR-126-5p	subtype	magenta
hsa-miR-126-5p	subtype	red
hsa-miR-1262	subtype	grey60
hsa-miR-127-3p	subtype	blue
hsa-miR-127-5p	subtype	magenta
hsa-miR-1271-5p	subtype	magenta
hsa-miR-1287-5p	subtype	blue
hsa-miR-1287-5p	subtype	brown
hsa-miR-1301-3p	subtype	pink
hsa-miR-1301-3p	subtype	grey60
hsa-miR-1304-3p	subtype	blue
hsa-miR-1304-5p	subtype	pink
hsa-miR-1306-5p	subtype	pink
hsa-miR-130a-3p	subtype	red
hsa-miR-130b-5p	subtype	blue
hsa-miR-136-5p	subtype	magenta
hsa-miR-136-5p	subtype	brown
hsa-miR-139-3p	subtype	magenta
hsa-miR-141-3p	subtype	grey60
hsa-miR-147b	subtype	grey60
hsa-miR-154-5p	subtype	magenta
hsa-miR-181d-5p	subtype	magenta
hsa-miR-182-5p	subtype	grey60
hsa-miR-183-5p	subtype	blue
hsa-miR-185-3p	subtype	pink
hsa-miR-18a-3p	subtype	pink
hsa-miR-193a-5p	subtype	brown
hsa-miR-195-3p	subtype	black
hsa-miR-195-5p	subtype	brown
hsa-miR-1976	subtype	grey60
hsa-miR-1976	subtype	pink
hsa-miR-199a-5p	subtype	magenta
hsa-miR-199a-5p	subtype	blue
hsa-miR-199b-5p	subtype	brown
hsa-miR-199b-5p	subtype	magenta
hsa-miR-200b-3p	subtype	grey60
hsa-miR-200c-3p	subtype	pink
hsa-miR-2110	subtype	pink
hsa-miR-2110	subtype	grey60
hsa-miR-214-5p	subtype	magenta
hsa-miR-218-5p	subtype	brown
hsa-miR-22-3p	subtype	brown
hsa-miR-2277-5p	subtype	blue
hsa-miR-23a-3p	subtype	magenta
hsa-miR-23a-5p	subtype	grey60
hsa-miR-23b-5p	subtype	blue
hsa-miR-23b-5p	subtype	pink
hsa-miR-301a-5p	subtype	blue
hsa-miR-3064-5p	subtype	blue
hsa-miR-3065-5p	subtype	pink
hsa-miR-30c-2-3p	subtype	blue
hsa-miR-30c-2-3p	subtype	black
hsa-miR-3158-3p	subtype	pink
hsa-miR-32-3p	subtype	black
hsa-miR-330-5p	subtype	grey60
hsa-miR-335-3p	subtype	magenta
hsa-miR-34b-3p	subtype	pink
hsa-miR-34b-5p	subtype	black
hsa-miR-34b-5p	subtype	grey60
hsa-miR-34c-5p	subtype	grey60
hsa-miR-3615	subtype	pink
hsa-miR-374b-3p	subtype	magenta
hsa-miR-376c-3p	subtype	magenta
hsa-miR-378a-3p	subtype	blue
hsa-miR-378c	subtype	grey60
hsa-miR-378g	subtype	magenta
hsa-miR-378g	subtype	blue
hsa-miR-379-3p	subtype	magenta
hsa-miR-3909	subtype	blue
hsa-miR-3909	subtype	brown
hsa-miR-3909	subtype	pink
hsa-miR-3913-5p	subtype	blue
hsa-miR-411-3p	subtype	magenta
hsa-miR-422a	subtype	blue
hsa-miR-423-5p	subtype	pink
hsa-miR-423-5p	subtype	grey60
hsa-miR-429	subtype	grey60
hsa-miR-4446-3p	subtype	black
hsa-miR-4446-3p	subtype	pink
hsa-miR-4446-3p	subtype	grey60
hsa-miR-455-5p	subtype	magenta
hsa-miR-4685-3p	subtype	grey60
hsa-miR-493-3p	subtype	magenta
hsa-miR-497-5p	subtype	magenta
hsa-miR-497-5p	subtype	brown
hsa-miR-5001-3p	subtype	pink
hsa-miR-504-5p	subtype	blue
hsa-miR-548e-3p	subtype	pink
hsa-miR-576-5p	subtype	black
hsa-miR-585-3p	subtype	magenta
hsa-miR-625-5p	subtype	pink
hsa-miR-629-5p	subtype	grey60
hsa-miR-642a-3p	subtype	grey60
hsa-miR-6504-5p	subtype	magenta
hsa-miR-6504-5p	subtype	brown
hsa-miR-654-3p	subtype	magenta
hsa-miR-655-3p	subtype	magenta
hsa-miR-656-3p	subtype	magenta
hsa-miR-671-3p	subtype	blue
hsa-miR-671-5p	subtype	grey60
hsa-miR-675-5p	subtype	grey60
hsa-miR-744-5p	subtype	grey60
hsa-miR-760	subtype	grey60
hsa-miR-7974	subtype	pink
hsa-miR-7974	subtype	magenta
hsa-miR-7974	subtype	brown
hsa-miR-92b-3p	subtype	pink
hsa-miR-92b-3p	subtype	black
hsa-miR-93-3p	subtype	pink
hsa-miR-941	subtype	blue
hsa-miR-942-5p	subtype	pink
hsa-miR-96-5p	subtype	grey60

Table 3.11 Functional enrichment of 14 coexpression modules discovered in lesion biopsies.

module	trait	module size	n module genes in term	Bonf. p-val	term ID	ontology	term name
black	grade, subtype	812	81	8.26E-84	GO:0022626	CC	cytosolic ribosome
black	grade, subtype	812	79	9.42E-80	GO:0006415	BP	translational termination
black	grade, subtype	812	86	2.04E-76	GO:0006414	BP	translational elongation
black	grade, subtype	812	81	1.26E-74	GO:0006614	BP	SRP-dependent cotranslational protein targeting to membrane
black	grade, subtype	812	81	1.86E-73	GO:0006613	BP	cotranslational protein targeting to membrane
black	grade, subtype	812	81	6.86E-73	GO:0045047	BP	protein targeting to ER
black	grade, subtype	812	81	6.86E-73	GO:0072599	BP	establishment of protein localization to endoplasmic reticulum
black	grade, subtype	812	81	1.04E-69	GO:0000184	BP	nuclear-transcribed mRNA catabolic process
black	grade, subtype	812	93	2.61E-68	GO:0006413	BP	translational initiation
black	grade, subtype	812	82	4.73E-67	GO:0070972	BP	protein localization to endoplasmic reticulum
blue	smoking, subtype	1755	521	2.33E-12	GO:0016070	BP	RNA metabolic process
blue	smoking, subtype	1755	419	1.69E-09	GO:0051252	BP	regulation of RNA metabolic process
blue	smoking, subtype	1755	463	2.15E-08	GO:0003676	MF	nucleic acid binding
blue	smoking, subtype	1755	413	2.31E-08	GO:0006351	BP	transcription
blue	smoking, subtype	1755	399	2.55E-07	GO:2001141	BP	regulation of RNA biosynthetic process
blue	smoking, subtype	1755	568	2.59E-07	GO:0010467	BP	gene expression
blue	smoking, subtype	1755	395	2.70E-07	GO:0006355	BP	regulation of transcription
blue	smoking, subtype	1755	293	2.84E-07	GO:0003677	MF	DNA binding
blue	smoking, subtype	1755	465	4.51E-07	GO:0010468	BP	regulation of gene expression
blue	smoking, subtype	1755	77	7.84E-07	GO:0008380	BP	RNA splicing
brown	grade, subtype	1141	203	0.001436157	GO:0003677	MF	DNA binding
brown	grade, subtype	1141	490	0.014226622	GO:0005634	CC	nucleus
brown	grade, subtype	1141	300	0.597424	GO:0003676	MF	nucleic acid binding
brown	grade, subtype	1141	8	1	GO:0010390	BP	histone monoubiquitination
brown	grade, subtype	1141	710	1	GO:0043231	CC	intracellular membrane-bounded organelle
brown	grade, subtype	1141	300	1	GO:0046872	MF	metal ion binding
brown	grade, subtype	1141	300	1	GO:0043169	MF	cation binding
brown	grade, subtype	1141	15	1	GO:0055072	BP	iron ion homeostasis
brown	grade, subtype	1141	40	1	GO:0000785	CC	chromatin
brown	grade, subtype	1141	71	1	GO:0005694	CC	chromosome
darkturquoise	grade	2412	289	1.80E-31	GO:0000278	BP	mitotic cell cycle
darkturquoise	grade	2412	534	1.23E-28	GO:0031981	CC	nuclear lumen
darkturquoise	grade	2412	394	3.22E-24	GO:0007049	BP	cell cycle
darkturquoise	grade	2412	311	4.95E-22	GO:0022402	BP	cell cycle process
darkturquoise	grade	2412	136	1.86E-19	GO:0007067	BP	mitotic nuclear division
darkturquoise	grade	2412	78	3.21E-17	GO:0007059	BP	chromosome segregation
darkturquoise	grade	2412	364	5.44E-17	GO:0005654	CC	nucleoplasm
darkturquoise	grade	2412	301	1.61E-15	GO:0044822	MF	poly(A) RNA binding
darkturquoise	grade	2412	73	3.02E-14	GO:0000793	CC	condensed chromosome
darkturquoise	grade	2412	192	2.45E-13	GO:0051301	BP	cell division
grey60	progr4, subtype, smoking	924	118	8.78E-59	GO:0031012	CC	extracellular matrix
grey60	progr4, subtype, smoking	924	109	6.09E-50	GO:0030198	BP	extracellular matrix organization
grey60	progr4, subtype, smoking	924	109	6.09E-50	GO:0043062	BP	extracellular structure organization
grey60	progr4, subtype, smoking	924	98	9.68E-50	GO:0005578	CC	proteinaceous extracellular matrix
grey60	progr4, subtype, smoking	924	408	1.50E-41	GO:0007275	BP	multicellular organismal development

module	trait	module size	n module genes in term	Bonf. p-val	termID	ontology	term name
grey60	progr4, subtype, smoking	924	172	8.80E-41	GO:0007155	BP	cell adhesion
grey60	progr4, subtype, smoking	924	446	1.06E-38	GO:0032502	BP	developmental process
grey60	progr4, subtype, smoking	924	382	1.20E-38	GO:0071944	CC	cell periphery
grey60	progr4, subtype, smoking	924	123	4.53E-38	GO:0001944	BP	vasculature development
grey60	progr4, subtype, smoking	924	373	1.25E-37	GO:0005886	CC	plasma membrane
lightgreen		875	372	8.69E-114	GO:0002376	BP	immune system process
lightgreen		875	270	2.74E-99	GO:0006955	BP	immune response
lightgreen		875	210	2.84E-88	GO:0001775	BP	cell activation
lightgreen		875	178	2.95E-83	GO:0045321	BP	leukocyte activation
lightgreen		875	224	2.03E-75	GO:0002682	BP	regulation of immune system process
lightgreen		875	241	3.08E-71	GO:0006952	BP	defense response
lightgreen		875	152	1.20E-69	GO:0046649	BP	lymphocyte activation
lightgreen		875	411	8.83E-64	GO:0071944	CC	cell periphery
lightgreen		875	404	1.84E-63	GO:0005886	CC	plasma membrane
lightgreen		875	170	6.23E-63	GO:0050776	BP	regulation of immune response
magenta	grade, subtype	3142	186	7.48E-56	GO:0005929	CC	cilium
magenta	grade, subtype	3142	107	1.55E-43	GO:0044782	BP	cilium organization
magenta	grade, subtype	3142	114	2.52E-43	GO:0060271	BP	cilium morphogenesis
magenta	grade, subtype	3142	96	2.00E-38	GO:0042384	BP	cilium assembly
magenta	grade, subtype	3142	50	6.13E-18	GO:0036064	CC	ciliary basal body
magenta	grade, subtype	3142	29	2.53E-17	GO:0030990	CC	intraciliary transport particle
magenta	grade, subtype	3142	44	6.97E-17	GO:0005930	CC	axoneme
magenta	grade, subtype	3142	46	4.58E-16	GO:0031514	CC	motile cilium
magenta	grade, subtype	3142	341	2.45E-15	GO:0042995	CC	cell projection
magenta	grade, subtype	3142	32	4.60E-15	GO:0003341	BP	cilium movement
pink	progr4, smoking, subtype	739	114	1.32E-13	GO:0048699	BP	generation of neurons
pink	progr4, smoking, subtype	739	107	2.54E-13	GO:0030182	BP	neuron differentiation
pink	progr4, smoking, subtype	739	117	9.08E-13	GO:0022008	BP	neurogenesis
pink	progr4, smoking, subtype	739	150	2.03E-12	GO:0007399	BP	nervous system development
pink	progr4, smoking, subtype	739	139	1.23E-10	GO:0048468	BP	cell development
pink	progr4, smoking, subtype	739	165	2.23E-09	GO:0009653	BP	anatomical structure morphogenesis
pink	progr4, smoking, subtype	739	257	4.30E-09	GO:0007275	BP	multicellular organismal development
pink	progr4, smoking, subtype	739	204	6.64E-09	GO:0030154	BP	cell differentiation
pink	progr4, smoking, subtype	739	78	1.40E-08	GO:0000904	BP	cell morphogenesis involved in differentiation
pink	progr4, smoking, subtype	739	78	1.79E-08	GO:0031175	BP	neuron projection development
red	smoking, subtype	943	158	4.01E-07	GO:0005739	CC	mitochondrion
red	smoking, subtype	943	86	1.75E-06	GO:0032446	BP	protein modification by small protein conjugation
red	smoking, subtype	943	81	4.03E-06	GO:0016567	BP	protein ubiquitination
red	smoking, subtype	943	20	6.32E-05	GO:0070469	CC	respiratory chain
red	smoking, subtype	943	7	8.76E-05	GO:0046977	MF	TAP binding
red	smoking, subtype	943	220	0.000159011	GO:0031090	CC	organelle membrane
red	smoking, subtype	943	60	0.000307633	GO:0019941	BP	modification-dependent protein catabolic process
red	smoking, subtype	943	636	0.000466202	GO:0043231	CC	intracellular membrane-bounded organelle
red	smoking, subtype	943	64	0.000552028	GO:0044257	BP	cellular protein catabolic process
red	smoking, subtype	943	7	0.000659805	GO:0019885	BP	antigen processing and presentation of endogenous peptide antigen via MHC class I

3.4. Discussion

PMLs are preinvasive histological abnormalities in the airway epithelium that can be reproducibly graded by a pathologist. Because many patients with lung cancer will also experience numerous severe PMLs, PMLs have been regarded as risk markers for any subtype of LC and the presumed precursors of SCC. Although the natural history of PMLs has been well documented, the proposed linear nature of lesion progression from hyperplasia to invasive carcinoma, does not reflect the often erratic behavior of lesions over time, as PMLs are dynamic and their histology may worsen and improve multiple times within a patient. Little is known about the molecular mechanisms behind PML progression. In fact, although we may be able to observe PMLs over time, it is currently not possible to predict if and when a lesion will progress to invasive carcinoma, relying solely on visual clues and patient's lesion history (histological grade, number and follow up status). In this chapter, we provide novel insights into transcriptional and regulatory processes likely governing PML progression by examining relevant expression changes demonstrated by genes and miRNAs in lesion biopsy samples over time.

We identified 31 miRNAs associated with PML histological grade or progression. Among those, we discovered four classes characterized by the co-association with traits of interest and control traits. While miRNAs contained in clusters 1, 3 and 4 associated with dysplasia grade, the biology of these miRNAs and mediating modules involved a wide array of mechanisms, including processes involved in cell proliferation,

differentiation, development and death– confirming that cancer may arise as a result of many disruptive cellular processes. In addition, miRNAs associated with multiple traits via multiple mediating modules could be of particular interest, although it is important to note that some of the associations may be driven by high correlation of genes belonging to multiple modules. Nonetheless, further exploration of miRNAs possibly governing multiple downstream biological processes, could yield additional insights into lesion progression.

Because a golden standard for defining lesion progression has not been established to date, we tested five definitions of progression, which reflect the potential capacity of the PMLs to store information about the past and the future, and the unknown rates at which gene and miRNA expression may respond to histological changes in PMLs over time. We observed gene and miRNA associations with the long-term future-based progression only. Since the *LT next* progression assumes comparison with the worst observed histology in the future, this is an encouraging result, as it suggests that the expression of genes and miRNAs observed in lesion biopsies might harbor information about PMLs being "preprogrammed" or destined to progress or regress. Although, all used samples represented a spectrum of premalignant disease and excluded invasive carcinoma, this fact would be especially useful in cases where extremely high or low expression of a progression-associated miRNA could be used to determine elevated risk for progressing into invasive carcinoma. Similarly, it may not be surprising to observe no associations with short-term progression, as the difference in time between consecutive

time points was not standardized and in many instances may be too short to be reflected on a molecular level (although, ST next was represented by “indirect” associations, but no miRNAs associated with it “directly” and thus there was no evidence for regulatory mechanisms implicated in this kind of progression). In addition, although it may be feasible for PMLs to harbor information about their worst histology observed in the past (much like crumpled paper that can never be brought back to the initial state, due to irreversible physical changes), past-based progression was not associated by any miRNAs and genes.

There are several limitations of our study. First, the relationship between a worsening histological state of individual PMLs and the overall health of the patient, and thus their LC risk, remains sparsely understood. Although we can define progression by comparing histology at consecutive time points, it is unclear if a single PML predicted to progress five grades (and potentially to invasive carcinoma), poses an equivalently elevated risk as multiple PMLs progressing in a milder manner (by three grades only for example). This means that while we could predict that a lesion will progress to carcinoma and direct a patient to a chemoprevention trial or administer a therapeutic, patients with multiple lesions progressing less aggressively might be at the same risk - we just don’t know the cumulative effect of PMLs on the overall health of the patient.

What is more, although we have observed significant associations with progression in our data, there may be other (perhaps better) definitions of progression we should consider, which may incorporate the number and severity of all PMLs present at a time. An

example would be progression defined based on a regression line fitted to the numerical histology data spaced proportionally according to time difference between time points (Figure 3.1. Example lung map of biopsy locations and corresponding histology grades changing over time.Figure 3.1). In theory, such regression line could provide a better overview of the total changes within a PML over time. However, we found that this measure was biased against subjects with large number of biopsies, for whom the slope of the regression line became closer to 0 with each additional time point (regression to the mean), and thus excluded this definition from consideration. Nonetheless, this proves that there are many different angles and perspectives for solving the problem of defining progression.

In addition, time difference between time points is not standardized, nor corrected for in this analysis. As discussed in Chapter 2, the individual follow-up histology of PMLs in the context of LT next progression may benefit from alternative definitions as well. Here, we consider the worst histology observed at later time points. However, one may argue that an average is more appropriate. There is a problem with both ideas - with enough time points, we might be able to observe severe dysplasia as worst histology, as well as metaplasia (the true transitional state) as the average in any PML.

Although we were able to identify miRNAs and genes whose expression patterns reflect progression in lesions, additional studies should be conducted to examine if any of these alterations are present in lesions that eventually progress into invasive carcinoma. However, these studies are difficult to conduct, as they require substantially long follow-

up time. For example, in a 2008 study by Salaun *et al* [104], 37 patients were monitored over the course of 12 years, which was sufficient to observe that 94% of lesions that progressed into invasive carcinoma were carcinomas *in situ* at baseline, and 79% of lesions that spontaneously regressed were severe dysplasias. After a similar follow up, a study by van Boerdonk *et al*. demonstrated that 55/164 patients with premalignant lesions developed lung cancer within that time frame (median time to event was 16.5 months) [136].

What is more, although a consensus coexpression network was built solely for the purpose of evaluating the reproducibility of the biopsy-derived modules, it will be important to explore the possibility of using the consensus network (instead of the reference network) for all analysis. Finally, additional evaluation of the results will be necessary in order to narrow down the list of grade - and progression-associated genes and miRNAs for experimental validation. This may include analyzing additional datasets and examining preservation of regulatory patterns, or selecting candidates confirmed individually in subjects with a high number of lesions to reduce the patient effect.

3.5. Conclusions

This chapter features a new paradigm for evaluating the role of miRNAs in biological processes involved in the progression of bronchial premalignant lesions. Leveraging the expression of genes likely targeted by miRNAs in our data, our results suggest that gene and miRNA expression extracted from bronchial premalignant lesion biopsies can be successfully used to determine PML severity and their capacity to

progress. The grade- and progression-associated genes were enriched in pathways often dysregulated in many cancers including that of the lung. In addition, many of the miRNAs associated with grade and progression play a significant role in inducing resistance to chemotherapy, promoting metastasis or tumor progression. This suggests that perturbation of these pathways in a precancerous environment may constitute a catalyst for future malignant changes, and that the commonalities between the regulatory processes involved in premalignancy and tumorigenesis may be the earliest events required for tumor development. In a clinical context, the presented methodology may provide a source of miRNA biomarkers that would facilitate identification of asymptomatic patients at high risk of developing lung cancer.

CHAPTER FOUR: Identifying Shared Genomic and Regulatory Alterations Associated with Lung Carcinogenesis in the Field, the Lesion and the Tumor

4.1. Background

Lung squamous cell premalignant lesions have been previously shown to be risk markers for developing lung cancer, as many patients with invasive carcinoma present with this histological abnormality of the airway [10,54]. However, despite the fact that the PMLs are the presumed precursors for SCC, their monitoring has not been widely adopted as part of lung cancer screening programs. In Chapter 2, we presented a gene expression-based biomarker derived from bronchial brushing samples that could be used to detect PML presence in high-risk individuals. In addition, in Chapter 3, we demonstrated that gene and miRNA expression extracted from lesion biopsies could be used to select a subset of subjects at highest risk for developing lung cancer due to the propensity of their lesions to progress, possibly to invasive carcinoma. Since the natural history of PMLs rarely follows the well-documented linear progression model and cannot be currently predicted by visual inspection and lesion history alone, the possibility of monitoring PMLs using gene and miRNA-expression opens the doors to new screening paradigms. However, due to the substantial invasiveness of procedures involving biopsy, in this chapter we sought to examine if bronchial brushings could be leveraged not only to detect PMLs but also to track the pathological changes associated with lesion progression over time. To address this challenge, we examined the preservation of

regulatory patterns characteristic of PML severity and progression we previously observed in lesion biopsies, in the context of airway field of injury.

4.2. Methods

4.2.1. Sample Collection: Bronchial Brushes from the PCGA

Bronchial airway brushes were obtained longitudinally from current and former smokers undergoing autofluorescence bronchoscopy at RPCI between December 2009 and March 2013 (For a detailed description of the RPCI cohort refer to Chapter 2). In addition, abnormally fluorescing areas (apparent and suspected PMLs) were sampled using endobronchial biopsy and subsequently graded by a pathologist. The worst observed PML histology was recorded and assigned to the bronchial brushing collected at the same time point.

4.2.2. Sample Collection: Tumor Biopsies from the TCGA

The TCGA cohort has previously been described in Chapter 2.

4.2.3. Redefining Lesion Grade and Progression

The definition of a lesion grade can no longer be directly applied to bronchial brushing samples, as they include only normal-appearing epithelial cells, while their pathologist-assigned histological grade reflects the worst histology observed among biopsied lesions. However, it seems reasonable to treat the brushing grade as equivalent to PML grade in all comparisons made in this Chapter.

Since no golden standard currently exists for defining PML progression, in Chapter 3 we propose 5 plausible definitions that account for PML's future and past histology modifications, as well as the varied time between procedures. Although the concept of progression is not as intuitive in the context of bronchial brushes, we classify brushing samples by applying the same definitions keeping in mind their more appropriate interpretation as progression of disease as opposed to progression of an individual lesion.

4.2.4. Identifying genes and miRNAs Associated with Grade and Progression in Bronchial Brushings

We applied miRCAT to bronchial brushing samples to identify genes and miRNAs whose expression was associated with grade or progression. We first built a brush-specific gene co-expression network using WGCNA and the same preprocessing steps and parameters as outlined for biopsies in Chapter 3. To facilitate a direct comparison between module assignments in biopsies and brushes, resulting modules were then tested for gene overlaps with the biopsy-derived modules, and those with significant overlaps were relabeled with an equivalent module color. The remaining steps of the miRCAT pipeline were then carried out without modification.

4.2.5. Testing Preservation of Biopsy-Derived Gene Modules Associated with PML Grade and Progression in Bronchial Brushes

Biopsy-derived co-expression module preservation in brushes

To evaluate if any of the grade- and progression-associated genes identified in lesion biopsies behaved similarly among the brushing samples, we employed the *modulePreservation* function in the WGCNA package. The brush-specific adjacency matrix as well as the biopsy-derived gene-module assignment were used as the test and reference networks respectively to calculate preservation statistics as described in [68]. Briefly, network attributes such as network density (average connection strength for all pairs of genes), and network-wide or intramodular node connectivity (or degree, the sum of connection strengths between a gene and all other genes in the network, or genes belonging to the same module, respectively), were calculated and combined into two composite preservation measures: (1) $Z_{summary}$ statistic, which sums the median values of multiple connectivity and density-based measures (making sure they are weighed equally) and classifies a module as strongly preserved if $Z_{summary} > 10$, weakly to moderately preserved if $2 < Z_{summary} < 10$, and not preserved if $Z_{summary} < 2$; as well as (2) a medianRank, which similarly allows for a direct comparison of multiple individual conservation statistics in multiple modules at once, but is less sensitive to module size due to its reliance on relative preservation among modules (as opposed to absolute measures).

Gene Set Enrichment Analysis

In addition, Gene Set Enrichment Analysis (GSEA) [126] was conducted to examine the enrichment of biopsy-derived grade and progression associated genes and miRNAs among brush-specific samples. Gene sets consisting of biopsy-derived module members were generated according to a t-value calculated for a linear mixed effects model correcting for batch and a random patient effect, modeling either dysplasia grade (treated as a numeric variable with values 1-6 corresponding to Normal – Severe dysplasia) or progression, by splitting them into submodules up- ($t>0$) and down-regulated ($t<0$) in the traits of interest. These gene sets were labeled with module color, trait and direction (e.g. magenta.grade.UP). Similar approach was applied to generate four miRNA sets, which involved grouping miRNAs by their associated trait and the directionality of that association. These gene sets were labeled with trait and direction of association (e.g. miRNA.progression.DOWN) In addition, target-specific gene sets were generated and included only neighboring target genes of each miRNA associated with a given trait in a given direction. These gene sets were labeled with the name of the miRNA, the trait and the direction (e.g. miR34b.grade.DOWN)

Ranked gene lists were generated in the brush data by ranking genes and miRNAs according to an LME t-value reflecting the strength and directionality of association with either grade or progression.

4.2.6. Testing Preservation of Biopsy-Derived miRNAs Associated with PML Grade and Progression in Lung SCC Tumors

GSEA was also conducted to examine the enrichment of biopsy-derived grade- and progression-associated genes and miRNAs among the TCGA lung SCC samples. The same gene sets were used as in the case of bronchial brushings. Ranked lists were generated in the TCGA data by ranking genes and miRNAs according to a Student's t statistic reflecting the strength and directionality of association with tissue type (tumor vs. adjacent normal).

4.3. Results

4.3.1. Sample Population: The PCGA brushes

86 brushing samples were collected longitudinally from high-risk current and former smokers 57 years of age on average, the majority of whom were Caucasian (93%) and female (57%). 54 brushings categorized as Dysplasia originated from subjects whose worst histological grade observed at that timepoint was mild dysplasia or worse. Conversely, 32 brushings categorized as Normal originated from subjects with the worst histology of Metaplasia or better (Table 4.1)

4.3.2. Sample Population: The TCGA

Sample population for the TCGA cohort has been previously described in Chapter 2.

4.3.3. Genes and miRNAs Associated with Traits of Interest in Bronchial Brushings

miRCAT pipeline was applied to 86 bronchial brushings to discover genes and miRNAs associated with grade, progression, or control traits such as smoking and subtype. We discovered 31 co-expression modules containing on average 485 genes, ranging in size between 42 and 2166 genes. Although we did not observe any universal relationships with any of the traits, 13 miRNAs exhibited direct association with smoking, and 6 miRNAs (3 overlapping) were associated indirectly with *short term previous* progression via the blue module (Table 4.2). let-7c-5p was the only progression-associated miRNA we also observed in biopsies.

4.3.4. Shared Gene and miRNA Alterations Present in the Field, the Lesion and the Tumor

Biopsy-derived gene co-expression modules were tested for preservation in the brushing data using network connectivity and density attributes. We found strong preservation of all modules ($Z_{summary} > 10$), with the exception of the progression-associated module grey60, which was moderately preserved ($Z_{summary} = 8.6$) (Figure 4.1).

In addition, we examined the enrichment of biopsy-derived genes and miRNAs associated with traits of interest in brushing samples. Four biopsy-derived modules mediating grade association (Table 3.7) were split into up- and downregulated submodules and utilized as gene sets in GSEA. The darkurquoiseUP, brownUP, and magentaDOWN were significantly enriched among the brushing data and tumor samples

in the corresponding direction (FDR q-val<0.05). The blackUP module was enriched specifically in tumors, while the blackDOWN and brownDOWN modules were only enriched among brushes. Interestingly, the black module was enriched oppositely in tumors as compared to brushes (Figure 4.2).

Two biopsy-derived modules mediating progression association (Table 3.8) were split into up- and down-regulated submodules and utilized as gene sets in GSEA. We did not find a significant (FDR q-val < 0.05) enrichment of the grey60 and pink submodules among the brushing data in either direction (Figure 4.2)

In addition to evaluating if the genes associated with grade and progression in biopsies behave similarly in brushes, we examined the preservation of expression patterns of miRNAs and their targets. Biopsy-derived miRNAs associated with grade and progression were split by trait into up- and down-regulated submodules and utilized as gene sets in GSEA. We found that only miRNAs up-regulated with progression were significantly enriched among upregulated brush miRNAs (FDR q=0.011) (Figure 4.3). All biopsy-derived miRNAs associated with progression were upregulated, thus we did not consider a gene set containing down-regulated miRNAs. When evaluating the preservation of miRNA targets however, we observed that 14 miRNA target sets associated with grade were enriched in brushes associated with dysplasia (FDR < 0.05; additional 6 were enriched with FDR<0.25). Two miRNA target sets upregulated in progression were enriched in brushes in the same direction (FDR<0.25). Finally, 16 miRNA target sets associated with grade were enriched in the corresponding direction

among tumor biopsies (FDR < 0.05; additional 4 were enriched with FDR<0.25) (Figure 4.4).

Table 4.1 Demographic characteristics of the RNA-Seq PML bronchial brushing dataset stratified by dysplasia status.
Data are means (SD) for continuous variables and counts for dichotomous variables. Top table summarizes data by samples, and bottom table summarizes data by subjects. P-values are for the comparison between the Dysplasia and Normal groups, using two sample t-tests for continuous variables or Fisher's exact test for categorical variables.

Factor	Overall	Dysplasia	Normal	p-value
No. Samples	86	54	32	
Age	57.89 (6.7)	57.95 (7.03)	57.79 (6.18)	0.91
Sex				0.26
female	49	28	21	
male	37	26	11	
Race				1
Caucasian	80	50	30	
Other	4	3	1	
Unknown	2	1	1	
Lung Cancer	0	0	0	
Smoking				*0.008*
current	40	31	9	
former	35	16	19	
NA	11	7	4	
Pack Years	49.35 (21.6)	48 (20.38)	51.62 (23.6)	0.46
Histological grade				*< 0.001*
Normal	6		6	
Hyperplasia	11		11	
Metaplasia	15		15	
Mild Dysplasia	8	8		
Moderate Dysplasia	9	9		
Severe Dysplasia	17	17		

Factor	Overall
No. Subjects	30
No. of procedures (mean [range])	2.9 (1-9 visits)
Days between visits (mean [range])	362 (84-1176 days)
Age	59.26 (7.45)
Sex	
female	15
male	15
Race	
Caucasian	27
African American	2
NA	1
Lung Cancer	0
Smoking	
current	12
former	12
NA	6
Pack Years	49.80 (22.11)

Table 4.2 Direct and indirect miRNA association with traits in brushes
miRNA association with traits including grade, progression, smoking and subtype, was evaluated using miRCAT pipeline. 13 miRNAs were associated directly via LME with smoking status. 6 miRNAs (3 overlapping) were associated indirectly with *short-term previous* progression via the blue module.

Direct association

miRbaseID	trait
hsa-let-7c-5p	smoking
hsa-miR-125a-5p	smoking
hsa-miR-125b-2-3p	smoking
hsa-miR-1295a	smoking
hsa-miR-190b	smoking
hsa-miR-2110	smoking
hsa-miR-30c-2-3p	smoking
hsa-miR-34c-3p	smoking
hsa-miR-34c-5p	smoking
hsa-miR-3607-3p	smoking
hsa-miR-642a-3p	smoking
hsa-miR-642a-5p	smoking
hsa-miR-99b-5p	smoking

Indirect association

miRbaseID	trait	module
hsa-miR-30e-3p	progr2	blue
hsa-miR-3182	progr2	blue
hsa-miR-3607-3p	progr2	blue
hsa-let-7f-5p	progr2	blue
hsa-let-7c-5p	progr2	blue
hsa-miR-125b-2-3p	progr2	blue

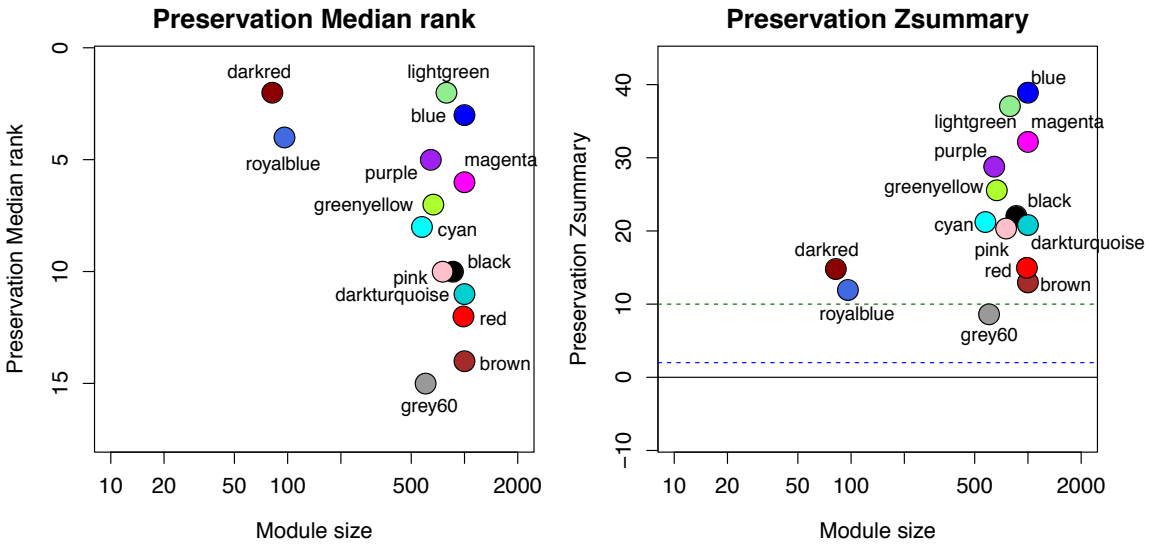


Figure 4.1 Biopsy-derived module preservation in brushing data.
Modules are considered strongly preserved if Zsummary>10, weakly preserved if 2<Zsummary<10, and not preserved if Zsummary<2. All modules are preserved strongly, with the exception of the gray60 module which was preserved weakly.

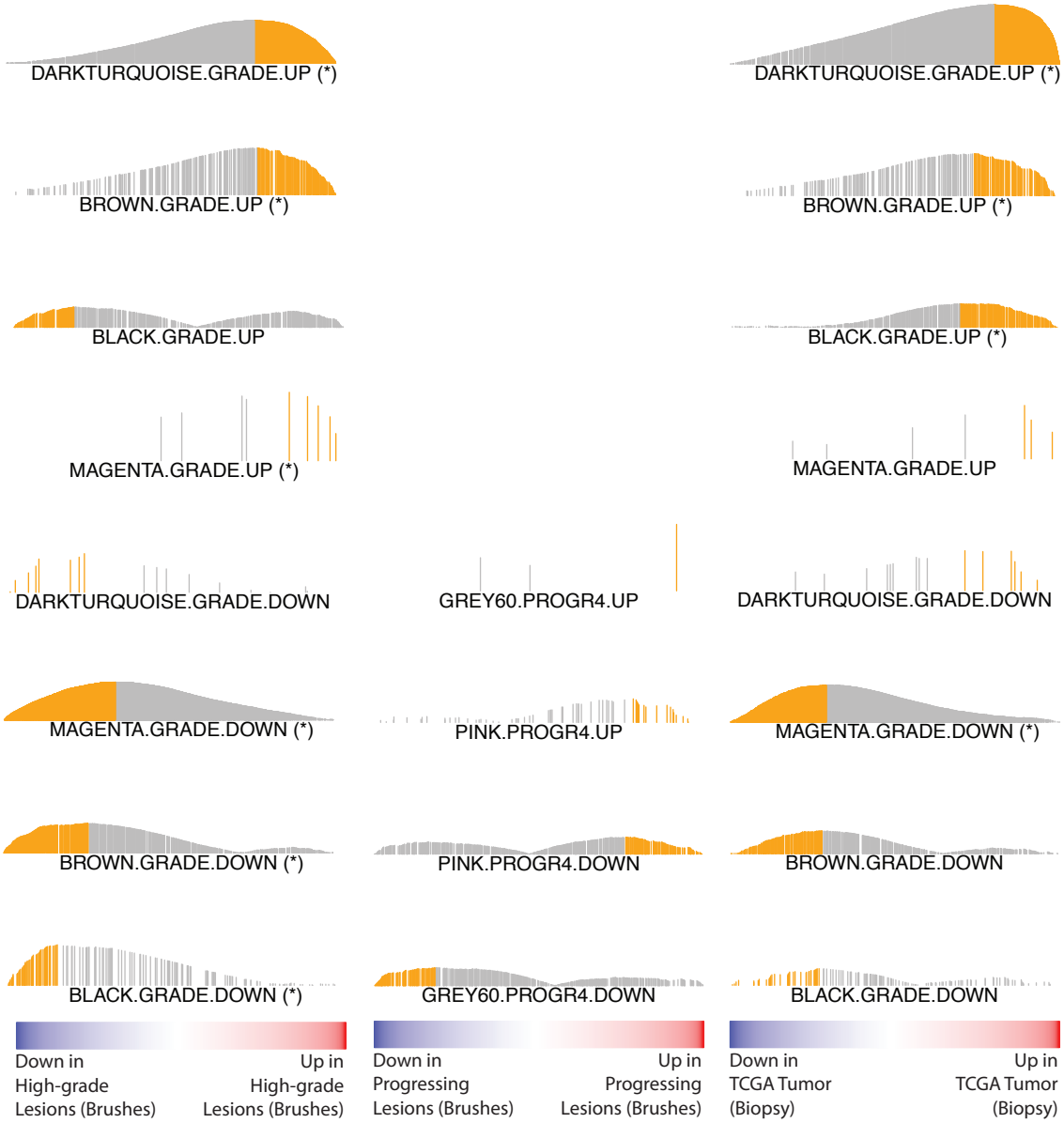


Figure 4.2 Biopsy-derived modules are enriched in brushes and tumor biopsies.
Gene sets with a significant enrichment score (FDR q-val < 0.05) are marked with (*). Grade-associated modules darkturquoiseUP, brownUP and magentaDOWN were significantly enriched in brushes and tumors. None of the progression-associated modules were enriched among brushing samples.

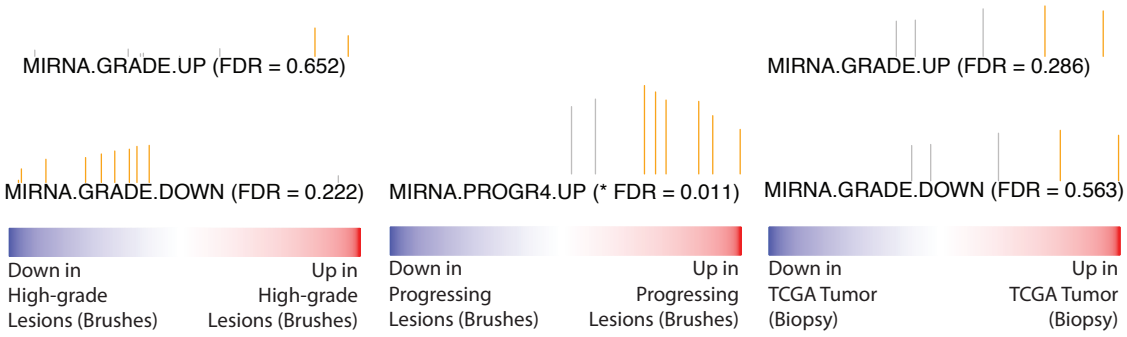


Figure 4.3 Trait associated miRNAs are enriched in brushes and tumor biopsies.
Gene sets with a significant enrichment score (FDR q-val < 0.05) are marked with (*). Only miRNAs positively associated with progression were enriched among up-regulated brushing samples.

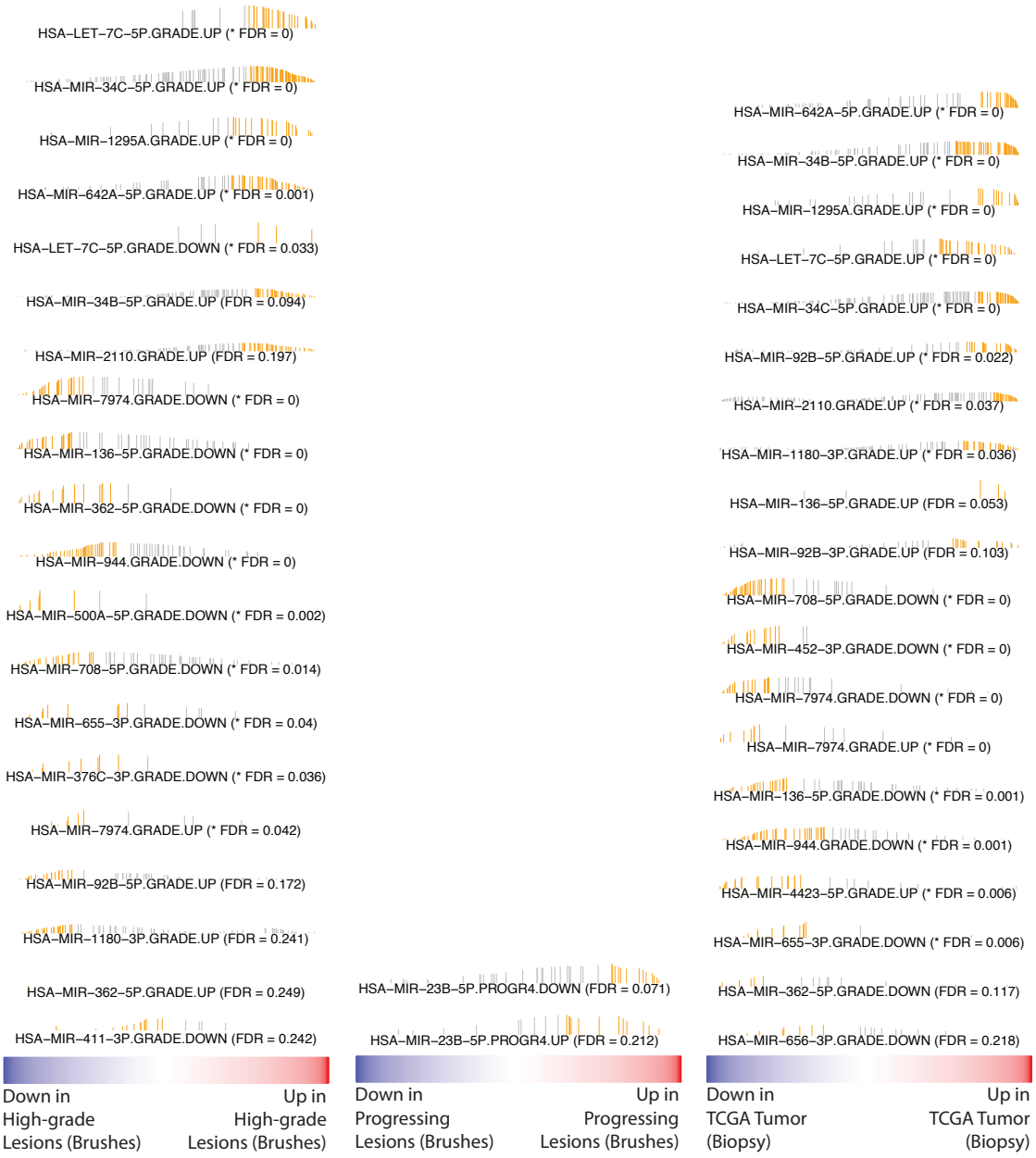


Figure 4.4 Targets of trait associated miRNAs are enriched in brushes and tumor biopsies. Gene sets with a significant enrichment score (FDR q-val < 0.05) are marked with (*). Only gene sets with FDR < 0.25 are presented. No targets of progression-associated miRNAs were significantly enriched among brushing samples

4.4. Discussion

PMLs have been previously shown to be risk markers for developing lung cancer. Although they are precursors for SCC, their monitoring has not been incorporated into lung cancer screening programs. In Chapter 2, we showed that we can leverage histologically normal brushing samples obtained non-invasively to detect the presence of PMLs. In Chapter 3, we demonstrated that gene and miRNA expression extracted from lesion biopsies facilitates the identification of regulatory patterns associated with lesion severity and progression. Since the natural history of PMLs tends to deviate from the well-established linear progression model and cannot be predicted by visual inspection, in this chapter we sought to examine if regulatory patterns similar to what we observed in biopsies existed in bronchial brushes. The presented findings suggest that gene and miRNA expression extracted from bronchial brushes could be used as a surrogate for the severity and potential for malignancy of precancerous lesions, and should be explored more extensively as a procedure facilitating PML monitoring without the necessity for invasive biopsy.

We first investigated if PML grade and progression signal could be extracted independently from bronchial brushings. Most likely due to the choice of parameters used in co-expression network construction that yielded approximately twice as many modules as in biopsies as well as a smaller sample size (in comparison to biopsies) that could affect correlation significance, we did not observe any universal associations between

miRNAs and traits of interest. However, several miRNAs were associated with *short term previous* progression indirectly. For example, miR-30e along with let-7e which were upregulated in progressing lesions, have been previously associated with decreased survival and dedifferentiation in NSCLC [12,152] – a process characteristic of cancer progression in which a differentiated cell might revert back to a less advanced type. miR-3182 has been shown to be upregulated in EGFR exon 19 deletion adenocarcinomas [58]. And the loss of miR-125b was reported to accompany melanoma onset and progression [59]. In addition, the association between these miRNAs and progression was mediated by a module implicated in transcription and regulation of RNA metabolic process, which has been described as one of the hallmarks of cancer [47]. These results suggest a potential for bronchial brushes to provide insights into PML development.

Moreover, we tested this hypothesis indirectly, by evaluating the preservation of biopsy-derived modules and miRNAs in brushes. We observed that independently of trait, all but one module were conserved between the two datasets. This gave us confidence that the relationships between genes in brushes has a high degree of similarity to the highly coexpressed gene modules we observed in biopsies. Next, we sought to evaluate enrichments of modules, miRNAs and their targets among brushes differentiated by PML grade or progression, as well as among lung tumor and adjacent normal samples. We found that three modules implicated in cell cycle (UP), ciliogenesis (DOWN) and nucleic acid binding (UP) pathways were enriched in the corresponding directions among brushes associated with grade as well as TCGA tumor vs. normal samples. The cilia (UP)

and nucleic acid binding (UP) submodules were also dysregulated in the same direction in brushes, but not in tumor. This might suggest processes specific to bronchial epithelial cells that are not as prominent in solid tumor tissues. On the other hand, a module implicated in translational elongation and termination (UP) was exclusively enriched among tumor samples. In the context of progression, no module gene sets were enriched among the progression-associated brushes. The lack of concordance between the gene sets is likely due to the fact that progression in biopsies describes a process that is lesion-specific, while grade changes in brushes over time reflect the changes in the worst observed histology in the biopsies. This was also confirmed by the finding that independently, miRNAs in brushes were associated (indirectly) with a different (past-based) definition of progression than biopsies. Surprisingly, when miRNAs were grouped by their association with grade and progression, the only enrichment we observed was of the upregulated progression-associated miRNAs in brushes upregulated with progression. However, the significance of enrichment could have been biased by the small size of the miRNA-set.

Finally, we evaluated the enrichment of biopsy-derived miRNAs in brushes and tumor biopsies on a miRNA target level. We observed concordant enrichment of targets of miRNAs associated with grade (UP and DOWN) in brushes as well as tumor biopsies (in the concordant direction), including let-7c, miR-34b/c, and miR-944.

There are several limitations to our study. As mentioned earlier, the lack of universal associations between miRNAs and grade and progression among bronchial

brushes is likely due to used parameters that were not tailored towards brush data in hopes of reducing the potential for technical biases that could influence the interpretation and comparison of biopsy and brush results. An additional experiment should be conducted in which module number is decreased. In addition, classifying brushing samples by the worst observed histology in biopsies is just one possible way of defining dysplasia grade and progression per patient/time point, and other definitions might be more appropriate. Finally, although module gene sets were split by the directionality of the association between member genes and traits of interest, it will be important to construct coexpression modules represented by genes either up- or down-regulated, but not both. This step would ensure that the up and down regulated module gene sets contained genes deemed coexpressed by the same approach instead of two (first, Pearson's correlation and average linkage hierarchical clustering to build modules, and then a linear mixed effects model to split the module into up and downregulated submodules). There is a risk that the resulting network will contain more, smaller modules, which had posed a bias in the brushing samples. However, the issue may become alleviated if we ensure singular-directionality of genes within. In addition, it may be interesting to explore the gene sets with bimodal enrichment score plots. For example, the brown.grade.DOWN module member-based gene set was significantly enriched in the up direction among high-grade brushing samples. However, it seems that a subset of gene set members occupying the non-leading edge switch directionality in the brushing dataset tested against. Also, among miRNA target-based gene sets, miR-1180 seems to be

oppositly enriched among the brushes (FDR < 0.25), but concordantly enriched among tumor biopsies (FDR < 0.05). Another kind of gene set could also be considered, which would allow for testing only unique miRNA-module-trait triplets that result from the indirect and direct association analysis. These new gene sets would include strongly correlated miRNA targets that belong to specific grade- or progression-associated modules, and would differ from the current target-based gene sets, by specifying separate gene sets for every mediating module. Specifically, since a miRNA can be associated with a trait via more than one module, currently, one target-based gene set is constructed to contain targets that belong to all mediating modules, while a separate new gene set would be constructed for each mediating module. This would ensure that biological processes that might not overlap between even highly correlated modules, are considered and interpreted separately. In the context of module preservation, tumor data could benefit from additional analyses, similar to what was done in brushes. Calculating Z-summary in the TCGA data would allow us to compare preservation between brushes and tumors.

4.5. Conclusions

In summary, the findings in this chapter lay the groundwork for leveraging cytologically-normal airway epithelial cell brushings to glean insights into the lung cancer premalignancy. While we did not observe independent brush-specific regulatory mechanisms associated with PMLs, we were able to demonstrate that a vast array of

miRNAs and their targets found to be associated with PML grade and progression in biopsy samples, behaved concordantly among bronchial brushings. Future studies will be necessary to reevaluate these commonalities, by either modifying existing methods or developing new ones. In addition, identifying a more appropriate definition for lesion grade and progression in the context of bronchial brushings that would capture the overall state of premalignancy in a patient (e.g. by incorporating multiplicity and precise locations of PMLs), will be crucial for proper interpretation of findings. As the link between PMLs and lung carcinogenesis becomes better understood, it will become important to consider PML monitoring as part of standard of care to aid in the identification of high-risk subjects. The possibility of capturing the changes in transcriptomic profiles from bronchial brushes, promises to decrease patient burden and provide wider accessibility by supplying a minimally invasive way to observe PMLs over time.

CHAPTER FIVE: General Conclusions and Future Directions

The studies featured in this dissertation collectively leverage the transcriptomic profiles extracted from bronchial brushes collected in the presence of premalignant lesions, and biopsies of the lesion itself to identify biological processes associated with PML presence, grade and progression. Importantly, the results from these chapters contend that:

- Cytologically-normal airway epithelial cells obtained via brushings of the main stem bronchus harbor biological information about the presence of PMLs within the central airway, and can be leveraged as a gene-expression based biomarker that can be administered using widely-available white-light bronchoscopy and bronchial brushing that is less invasive than biopsy. Ultimately, the biomarker could achieve the greatest clinical utility if the sampling was moved to a non-invasive location, such as the nasal epithelium, where high-risk subjects could be identified more readily even if they didn't meet the enrollment criteria for existing screening programs.
- PMLs, which are the presumed precursors of lung squamous cell carcinoma, are risk markers for developing lung cancer. Although PMLs are thought to follow a linear progression model, changes in their histology over time are rarely predictable by inspection of visual clues or lesion's history. However, the severity and progression of PMLs is reflected in the gene and miRNA expression profiles extracted from bronchial premalignant lesion biopsies, monitoring of which could

(a) aid in the identification of asymptomatic patients with high-grade lesions likely to progress who need a more aggressive follow-up and (b) help facilitate disease interception. In addition, the miRCAT approach, which constitutes a new paradigm for evaluating the role of miRNAs in biological processes involved in the progression of bronchial premalignant lesions, promises great utility in the identification of biomarkers of progression and response to therapy, as well as the selection of appropriate therapeutic targets

- The airway field of injury recapitulates a subset of grade and progression-associated alterations observed in lesions. Although PMLs are typically detected using autofluorescence bronchoscopy, the procedure's scarcity makes it unrealistic to be widely used in screening programs. Building on our findings that demonstrate that bronchial brushes obtained via white-light bronchoscopy can successfully identify individuals with PMLs, and that lesion biopsies carry information that can be harnessed to identify subjects with lesions likely to progress, we proposed an alternative to biopsy's invasiveness: Capturing the changes in transcriptomic profiles from bronchial brushes, promises to provide wider accessibility and decrease patient burden by supplying a minimally invasive way to observe PMLs over time.

BIBLIOGRAPHY

1. Allen TC. Pulmonary Neoplasia. October 2009. doi:10.1043/1543-2165(2008)132[1053:PN]2.0.CO;2.

2. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome biology*. 2010;11(10):1. doi:10.1186/gb-2010-11-10-r106.

3. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics (Oxford, England)*. 2015;31(2):166-169. doi:10.1093/bioinformatics/btu638.

4. Anders S, Pyl PT, Huber W. HTSeq–A Python framework to work with high-throughput sequencing data. *bioRxiv*. 2014. doi:10.1101/002824.

5. Andrés-León E, Peña DG, Gómez-López G, Pisano DG. miRGate: a curated database of human, mouse and rat miRNA–mRNA targets. *Database*. 2015;2015(0):bav035-bav035. doi:10.1093/database/bav035.

6. Androutsopoulos VP, Tsatsakis AM, Spandidos DA. Cytochrome P450 CYP1A1: wider roles in cancer progression and prevention. *BMC Cancer*. 2009;9(1):3975. doi:10.1186/1471-2407-9-187.

7. Atmaca A, Al-Batran S-E, Wirtz RM, et al. The validation of estrogen receptor 1 mRNA expression as a predictor of outcome in patients with metastatic non-small cell lung cancer. - PubMed - NCBI. *International Journal of Cancer*. 2013;134(10):2314-2321. doi:10.1002/ijc.28571.

8. AUERBACH O, FORMAN JB, GERE JB, et al. Changes in the bronchial epithelium in relation to smoking and cancer of the lung; a report of progress. *New England Journal of Medicine*. 1957;256(3):97-104. doi:10.1056/NEJM195701172560301.

9. BALÇA-SILVA J, NEVES SS, GONÇALVES AC, et al. Effect of miR-34b Overexpression on the Radiosensitivity of Non-small Cell Lung Cancer Cell Lines. *Anticancer Research*. 2012;32(5):1603-1609. doi:10.1016/j.molcel.2007.05.017.

10. Banerjee AK. Preinvasive lesions of the bronchus. *Journal of Thoracic Oncology*. 2009;4(4):545-551. doi:10.1097/JTO.0b013e31819667bd.

11. Barabási A-L, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. *Nature Reviews. Genetics*. 2011;12(1):56-68.

doi:10.1038/nrg2918.

12. Barh D, Malhotra R, Ravi B, Sindhurani P. MicroRNA let-7: an emerging next-generation cancer therapeutic. *Current Oncology*. 2010;17(1):70-80.

13. Bates D, Mächler M, Ben Bolker, Walker S. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*. 2015;67(1):1-48. doi:10.18637/jss.v067.i01.

14. Beane J, Beane J, Sebastiani P, et al. A Prediction Model for Lung Cancer Diagnosis that Integrates Genomic and Clinical Features. *Cancer Prevention Research.* 2008;1(1):56-64. doi:10.1158/1940-6207.CAPR-08-0011.

15. Beane J, Sebastiani P, Liu G, Brody JS, Lenburg ME, Spira A. Reversible and permanent effects of tobacco smoke exposure on airway epithelial gene expression. *Genome Biology*. 2007;8(9):R201. doi:10.1186/gb-2007-8-9-r201.

16. Beane J, Sebastiani P, Liu G, Brody JS, Lenburg ME, Spira A. Reversible and permanent effects of tobacco smoke exposure on airway epithelial gene expression. *Genome Biology*. 2007;8(9):R201. doi:10.1186/gb-2007-8-9-r201.

17. Belinsky SA, Nikula KJ, Palmisano WA, et al. Aberrant methylation of p16(INK4a) is an early event in lung cancer and a potential biomarker for early diagnosis. *Proceedings of the National Academy of Sciences of the United States of America*. 1998;95(20):11891-11896.

18. Bin Zhang, Gaiteri C, Bodea L-G, et al. Integrated Systems Approach Identifies Genetic Nodes and Networks in Late-Onset Alzheimer's Disease. *Cell*. 2013;153(3):707-720. doi:10.1016/j.cell.2013.03.030.

19. Blomquist T, Crawford EL, Mullins D, et al. Pattern of antioxidant and DNA repair gene expression in normal airway epithelium associated with lung cancer diagnosis. *Cancer Research*. 2009;69(22):8629-8635. doi:10.1158/0008-5472.CAN-09-1568.

20. Blum R, Kloog Y. Metabolism addiction in pancreatic cancer. *Cell Death & Disease*. 2014;5(2):e1065. doi:10.1038/cddis.2014.38.

21. Boeri M, Verri C, Conte D, et al. MicroRNA signatures in tissues and plasma predict development and prognosis of computed tomography detected lung cancer. *Proceedings of the National Academy of Sciences*.

2011;108(9):3713-3718. doi:10.1073/pnas.1100048108.

22. Breuer RH, Pasic A, Smit EF, et al. The Natural Course of Preneoplastic Lesions in Bronchial Epithelium. *Clinical Cancer Research.* 2005;11(2):537-543. doi:10.1183/09031936.01.00275301.

23. Brothers JF, Hijazi K, Mascaux C, El-Zein RA, Spitz MR, Spira A. Bridging the clinical gaps: genetic, epigenetic and transcriptomic biomarkers for the early detection of lung cancer in the post-National Lung Screening Trial era. *BMC Medicine 2013 11:1*. 2013;11(1):168. doi:10.1186/1741-7015-11-168.

24. Buja A, Eyuboglu N. Remarks on Parallel Analysis. *Multivariate Behavioral Research*. 1992;27(4):509-540. doi:10.1207/s15327906mbr2704_2.

25. Campbell JD, Campbell JD, Mazzilli SA, et al. The Case for a Pre-Cancer Genome Atlas (PCGA). *Cancer Prevention Research*. 2016;9(2):119-124. doi:10.1158/1940-6207.CAPR-16-0024.

26. Campbell JD, Liu G, Luo L, et al. Assessment of microRNA differential expression and detection in multiplexed small RNA sequencing data. *RNA*. 2015;21(2):164-171. doi:10.1261/rna.046060.114.

27. Carter AJ, Nguyen CN. A comparison of cancer burden and research spending reveals discrepancies in the distribution of research funding. *BMC Public Health*. 2012;12(1):277. doi:10.1186/1471-2458-12-526.

28. Chen EY, Tan CM, Kou Y, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*. 2013;14:128-128. doi:10.1186/1471-2105-14-128.

29. Chimal JB. Hsp72 is an early and sensitive biomarker to detect acute kidney injury. *Nat Rev Nephrol*. 2010;6(2):71-73. doi:10.1038/nrneph.2009.225.

30. Cortes C, Vapnik V. Support-vector networks. *Machine Learning*. 1995; 20(3):273-297. doi:10.1007/BF00994018.

31. Cortese DA, Pairolero PC, Bergstralh EJ, et al. Roentgenographically occult lung cancer. A ten-year experience. *Journal of Thoracic and Cardiovascular Surgery.* 1983;86(3):373-380.

32. Detterbeck FC. Overdiagnosis during lung cancer screening: is it an overemphasised, underappreciated, or tangential issue? *Thorax*. 2014;69(5):407-408. doi:10.1136/thoraxjnl-2014-205140.

33. Diaz LA, Bardelli A. Liquid Biopsies: Genotyping Circulating Tumor DNA: Journal of Clinical Oncology: Vol 32, No 6. *Journal of Clinical Oncology*. 2014.

34. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2012;29(1):bts635–21. doi:10.1093/bioinformatics/bts635.

35. Esteller M, Sanchez-Cespedes M, Rosell R, Sidransky D, Baylin SB, Herman JG. Detection of aberrant promoter hypermethylation of tumor suppressor genes in serum DNA from non-small cell lung cancer patients. *Cancer research*. 1999;59(1):67-70.

36. Feng J, Wang Y, Qing, Qi. Comparison of autofluorescence imaging bronchoscopy and white light bronchoscopy for detection of lung cancers and precancerous lesions. *PPA*. 2013;7:621-631. doi:10.2147/PPA.S46749.

37. Forshew T, Murtaza M, Parkinson C, et al. Noninvasive Identification and Monitoring of Cancer Mutations by Targeted Deep Sequencing of Plasma DNA. *Science translational medicine*. 2012;4(136):136ra68-136ra68. doi:10.1126/scitranslmed.3003726.

38. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*. 2010;33(1):1-968. doi:10.1109/TPAMI.2005.127.

39. Gentleman R, Carey V, Huber W, Hahne F. *Genefilter: Methods for Filtering Genes From High-Throughput Experiments*. R package version; 2015.

40. Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*. 1999;286(5439):531-537.

41. Goossens N, Nakagawa S, Sun X, Hoshida Y. Cancer biomarker discovery and validation. *Translational Cancer Research*. 2015;4(3):256-269. doi:10.3978/j.issn.2218-676X.2015.06.04.

42. Griffiths Jones S. The microRNA Registry. *Nucleic Acids Research*. 2004;32(suppl_1):D109-D111. doi:10.1093/nar/gkh023.

43. Guan G, Zhang D, Zheng Y, et al. microRNA-423-3p promotes tumor progression via modulation of AdipoR2 in laryngeal carcinoma. *International*

Journal of Clinical and Experimental Pathology. 2014;7(9):5683-5691.

44. Gustafson AM, Soldi R, Anderlind C, et al. Airway PI3K pathway activation is an early and reversible event in lung cancer development. *Science Translational Medicine*. 2010;2(26):26ra25-26ra25. doi:10.1126/scitranslmed.3000251.

45. Ha M, Kim VN. Regulation of microRNA biogenesis. *Nature Reviews. Molecular Cell Biology*. 2014;15(8):509-524. doi:10.1038/nrm3838.

46. Hackett NR, Heguy A, Harvey B-G, et al. Variability of Antioxidant-Related Gene Expression in the Airway Epithelium of Cigarette Smokers. *Am J Respir Cell Mol Biol*. 2003;29(3):331-343. doi:10.1165/rcmb.2002-0321OC.

47. Hanahan D, Weinberg RA. Hallmarks of Cancer: The Next Generation. *Cell*. 2011;144(5):646-674. doi:10.1016/j.cell.2011.02.013.

48. Häussinger K, Becker H, Stanzel F, et al. Autofluorescence bronchoscopy with white light bronchoscopy compared with white light bronchoscopy alone for the detection of precancerous lesions: a European randomised controlled multicentre trial. *Thorax*. 2005;60(6):496-503. doi:10.1136/thx.2005.041475.

49. Hermeking H. The miR-34 family in cancer and apoptosis. *Cell Death & Differentiation*. 2010;17(2):193-199. doi:10.1038/cdd.2009.56.

50. Hoffmann D, Hoffmann I. The less harmful cigarette: a controversial issue. A tribute to Ernst L. Wynder. *Chemical Research in Toxicology*. 2001:14(7): 767-790. doi:10.1021/tx000260u.

51. Howlader N, AM N, M K, et al. SEER Cancer Statistics Review, 1975-2013. http://seer.cancer.gov/csr/1975_2013/. http://seer.cancer.gov/csr/1975_2013/. Published 2016. Accessed March 25, 2017.

52. Hubers AJ, Heideman DAM, Burgers SA, et al. DNA hypermethylation analysis in sputum for the diagnosis of lung cancer: training validation set approach. *Brit J Cancer*. 2015;112(6):1105-1113. doi:10.1038/bjc.2014.636.

53. Irizarry RA, Hobbs B, Collin F, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics (Oxford, England)*. 2003;4(2):249-264. doi:10.1093/biostatistics/4.2.249.

54. Ishizumi T, McWilliams A, MacAulay C, Gazdar A, Lam S. Natural history of bronchial preinvasive lesions. *Cancer Metastasis Reviews*. 2010;29(1):5-14. doi:10.1007/s10555-010-9214-7.

55. Ishizumi T, McWilliams A, MacAulay C, Gazdar A, Lam S. Natural history of bronchial preinvasive lesions. *Cancer Metastasis Rev*. 2010;29(1):5-14. doi:10.1007/s10555-010-9214-7.

56. Jordan JD, Landau EM, Iyengar R. Signaling networks: the origins of cellular multitasking. *Cell*. 2000;103(2):193-200. doi:10.1016/S0092-8674(00)00112-4.

57. Jovanovic M, Hengartner MO. miRNAs and apoptosis: RNAs to die for. *Oncogene*. 2006;25(46):6176-6187. doi:10.1038/sj.onc.1209912.

58. Ju L, Han M, Zhao C, Li X. Genome-wide analysis of microRNA signature in lung adenocarcinoma with EGFR exon 19 deletion. *bioRxiv*. 2016. doi:10.1101/032367.

59. Kappelmann M, Kuphal S, Meister G, Vardimon L, Bosserhoff A-K. MicroRNA miR-125b controls melanoma progression by direct regulation of c-Jun protein expression. *Oncogene*. 2013;32(24):2984-2991. doi:10.1038/onc.2012.307.

60. Karlebach G, Shamir R. Modelling and analysis of gene regulatory networks. *Nature Reviews. Molecular Cell Biology*. 2008;9(10):770-780. doi:10.1038/nrm2503.

61. Katz Y, Wang ET, Airoldi EM, Burge CB. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature Methods*. 2010;7(12):1009-1015. doi:10.1038/nmeth.1528.

62. Keith RL, Blatchford PJ, Kittelson J, et al. Oral Iloprost Improves Endobronchial Dysplasia in Former Smokers. *Cancer Prevention Research*. 2011;4(6):793-802. doi:10.1158/1940-6207.CAPR-11-0057.

63. Keith RL, Miller YE. Lung cancer chemoprevention: current status and future prospects. *Nature Reviews. Clinical Oncology*. 2013;10(6):334-343. doi:10.1038/nrclinonc.2013.64.

64. Kim K-H, Cho E-G, Yu SJ, et al. ΔNp63 intronic miR-944 is implicated in the ΔNp63-mediated induction of epidermal differentiation. *Nucleic Acids Research*. 2015;43(15):7462-7479. doi:10.1093/nar/gkv735.

65. Kneip C, Schmidt B, Seegebarth A, et al. SHOX2 DNA Methylation Is a Biomarker for the Diagnosis of Lung Cancer in Plasma. *Journal of Thoracic Oncology*. 2011;6(10):1632-38. doi:10.1097/JTO.0b013e318220ef9a.

66. Kukurba KR, Montgomery SB. RNA Sequencing and Analysis. *Cold Spring Harbor Protoc*. 2015;2015(11):pdb.top084970. doi:10.1101/pdb.top084970.

67. Lam S, LeRiche JC, Zheng Y, et al. Sex-related differences in bronchial epithelial changes associated with tobacco smoking. *Journal of the National Cancer Institute.* 1999;91(8):691-696.

68. Langfelder P, Luo R, Oldham MC, Horvath S. Is my network module preserved and reproducible? *PLoS Computational Biology*. 2011;7(1):e1001057. doi:10.1371/journal.pcbi.1001057.

69. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*. 2009;10(3):R25. doi:10.1186/gb-2009-10-3-r25.

70. Law CW, Chen Y, Shi W, Smyth GK. Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol*ogy. 2014.

71. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics (Oxford, England)*. 2012;28(6):882-883. doi:10.1093/bioinformatics/bts034.

72. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*. 2011;12(1):323. doi:10.1186/1471-2105-12-323.

73. Liaw A, Wiener M. Classification and regression by randomForest. *R news*. 2002.

74. Lockwood WW, Wilson IM, Coe BP, et al. Divergent Genomic and Epigenomic Landscapes of Lung Cancer Subtypes Underscore the Selection of Different Oncogenic Pathways during Tumor Development. Navarro A, ed. *PloS one*. 2012;7(5):e37775. doi:10.1371/journal.pone.0037775.

75. MacFarlane L-A, R Murphy P. MicroRNA: Biogenesis, Function and Role in Cancer. *CG*. 2010;11(7):537-561. doi:10.2174/138920210793175895.

76. Mapstone M, Cheema AK, Fiandaca MS, et al. Plasma phospholipids

identify antecedent memory impairment in older adults. *Nature Medicine*. 2014;20(4):415-418. doi:10.1038/nm.3466.

77. Marchetti A, Palma JF, Felicioni L, et al. Early Prediction of Response to Tyrosine Kinase Inhibitors by Quantification of EGFR Mutations in Plasma of NSCLC Patients. *Journal of Thoracic Oncology.* 2015;10(10):1437-1443. doi:10.1097/JTO.0000000000000643.

78. Martinez-Outschoorn UE, Pavlides S, Sotgia F, Lisanti MP. Mitochondrial Biogenesis Drives Tumor Cell Proliferation. *The American Journal of Pathology*. 2011;178(5):1949-1952. doi:10.1016/j.ajpath.2011.03.002.

79. Mascaux C, Laes JF, Anthoine G, et al. Evolution of microRNA expression during human bronchial squamous carcinogenesis. *European Respiratory Journal.* 2008;33(2):352-359. doi:10.1183/09031936.00084108.

80. McClish DK. Analyzing a portion of the ROC curve. *Medical Decision Making.* 1989;9(3):190-195.

81. McGhee D, Royle PL, Thompson PA. A systematic review of biomarkers for disease progression in Parkinson's disease. *BMC Neurology*. 2013;13:35.

82. McGhee DJM, Ritchie CW, Thompson PA, Wright DE, Zajicek JP, Counsell CE. A Systematic Review of Biomarkers for Disease Progression in Alzheimer's Disease. *PloS One*. 2014;9(2):e88854. doi:10.1371/journal.pone.0088854.

83. Merrick DT, Gao D, Miller YE, et al. Persistence of Bronchial Dysplasia Is Associated with Development of Invasive Squamous Cell Carcinoma. *Cancer Prevention Research*. 2016;9(1):96-104. doi:10.1158/1940-6207. CAPR-15-0305.

84. Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F. *E1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien [R Package E1071 Version 1.6-7]*. 1st ed. Comprehensive R Archive Network (CRAN); 2015. http://CRAN.R-project.org/package=e1071.

85. Navarro F, Lieberman J. miR-34 and p53: New Insights into a Complex Functional Relationship. Martelli F, ed. *PloS One*. 2015;10(7):e0132767. doi:10.1371/journal.pone.0132767.

86. Nicholson AG, Perry LJ, Cury PM, et al. Reproducibility of the

WHO/IASLC grading system for pre-invasive squamous lesions of the bronchus: a study of inter-observer and intra-observer variation. *Histopathology*. 2001;38(3):202-208.

87. Nygaard AD, Garm Spindler K-L, Pallisgaard N, Andersen RF, Jakobsen A. The prognostic value of KRAS mutated plasma DNA in advanced non-small cell lung cancer. *Lung Cancer*. 2013;79(3):312-317. doi:10.1016/j.lungcan.2012.11.016.

88. Overington JP, Al-Lazikani B, Hopkins AL. How many drug targets are there? *Nat Reviews. Drug Discov*. 2006;5(12):993-96. doi:10.1038/nrd2199.

89. Oxnard GR, Paweletz CP, Kuang Y, et al. Noninvasive Detection of Response and Resistance in EGFR-Mutant Lung Cancer Using Quantitative Next-Generation Genotyping of Cell-Free Plasma DNA. *Clinical Cancer Research.* 2014;20(6):1698-1705. doi:10.1158/1078-0432.CCR-13-2482.

90. Paris C, Benichou J, Bota S, et al. Occupational and nonoccupational factors associated with high grade bronchial pre-invasive lesions. *European Respiratory Journal.* 2003;21(2):332-341.

91. Patz EF, Pinsky P, Gatsonis C, et al. Overdiagnosis in low-dose computed tomography screening for lung cancer. *JAMA Internal Medicine*. 2014; 174(2):269-274. doi:10.1001/jamainternmed.2013.12738.

92. Pearson K. LIII. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine Series 6*. 1901;2(11):559-572. doi:10.1080/14786440109462720.

93. Peng Y, Croce CM. The role of MicroRNAs in human cancer. *Signal Transduction and Targeted Therapy*. 2016;1:15004. doi:10.1038/sigtrans.2015.4.

94. Perdomo C, Campbell JD, Gerrein J, et al. MicroRNA 4423 is a primate-specific regulator of airway epithelial cell differentiation and lung carcinogenesis. *Proceedings of the National Academy of Sciences U S A*. 2013;110(47):18946-18951. doi:10.1073/pnas.1220319110.

95. Piccolo SR, Sun Y, Campbell JD, Lenburg ME, Bild AH, Johnson WE. A single-sample microarray normalization method to facilitate personalized-medicine workflows. *Genomics*. 2012;100(6):337-344. doi:10.1016/j.ygeno.2012.08.003.

96. Pierotti MA, Sozzi G, Croce CM. Mechanisms of oncogene activation. 2003. https://www.ncbi.nlm.nih.gov/books/NBK12538/.

97. Polakis P. Wnt Signaling in Cancer. *Cold Spring Harbor Perspectives in Biology*. 2012;4(5):a008052-a008052. doi:10.1101/cshperspect.a008052.

98. Powrózek T, Krawczyk P, Kowalski DM, Winiarczyk K, Olszyna-Serementa M, Milanowski J. Plasma circulating microRNA-944 and microRNA-3662 as potential histologic type-specific early lung cancer biomarkers. *Translational Research*. 2015;166(4):315-323. doi:10.1016/j.trsl.2015.05.009.

99. Ritchie ME, Phipson B, Di Wu, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*. 2015;43(7):gkv007–e47. doi:10.1093/nar/gkv007.

100. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*. 2011;12(1):77. doi:10.1186/1471-2105-12-77.

101. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2010;26(1):139-140. doi:10.1093/bioinformatics/btp616.

102. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*. 2010;11(3):R25. doi:10.1186/gb-2010-11-3-r25.

103. Saccomanno G. Carcinoma in-situ of the lung: Its development, detection, and treatment. *Seminars in respiratory medicine*. 1982. doi:10.1055/s-2007-1012480.pdf.

104. Salaün M, Sesboüé R, Moreno-Swirc S, et al. Molecular Predictive Factors for Progression of High-Grade Preinvasive Bronchial Lesions. *Am J Respir Crit Care Med*. 2008;177(8):880-886. doi:10.1164/rccm.200704-598OC.

105. Schembri F, Sridhar S, Perdomo C, et al. MicroRNAs as modulators of smoking-induced gene expression changes in human airway epithelium. *Proceedings of the National Academy of Sciences*. 2009;106(7):2319-2324. doi:10.1073/pnas.0806383106.

106. Schmidt B, Liebenberg V, Dietrich D, et al. SHOX2 DNA Methylation is a Biomarker for the diagnosis of lung cancer based on bronchial aspirates. *BMC Cancer*. 2010;10(1):71. doi:10.1186/1471-2407-10-600.

107. Segrè D, DeLuna A, Church GM, Kishony R. Modular epistasis in yeast metabolism. - PubMed - NCBI. *Nature Genetics*. 2004;49(1):703-783. doi:10.1038/ng1489.

108. Seguro AC, ed. Hsp72 Is a Novel Biomarker to Predict Acute Kidney Injury in Critically Ill Patients. *PloS One*. 2014;9(10):e109407. doi:10.1371/journal.pone.0109407.

109. Sestini P. Reduced lung-cancer mortality with CT screening. *N Engl J Med*. 2011;365(21):2037–authorreply2037–8. doi:10.1056/NEJMc1110293#SA5.

110. Sestini S, Boeri M, Marchiano A, et al. Circulating microRNA signature as liquid-biopsy to monitor lung cancer in low-dose computed tomography screening. *Oncotarget*. 2015;6(32):32868-32877. doi:10.18632/oncotarget.5210.

111. Shah V, Sridhar S, Beane J, Brody JS, Spira A. SIEGE: Smoking Induced Epithelial Gene Expression Database. *Nucleic Acids Research*. 2005;33 (Database issue):D573-D579. doi:10.1093/nar/gki035.

112. Shen J, Liu Z, Todd NW, et al. Diagnosis of lung cancer in individuals with solitary pulmonary nodules by plasma microRNA biomarkers. *BMC Cancer*. 2011;11(1):71. doi:10.1186/1471-2407-11-374.

113. Showe MK, Vachani A, Kossenkov AV, et al. Gene Expression Profiles in Peripheral Blood Mononuclear Cells Can Distinguish Patients with Non-Small Cell Lung Cancer from Patients with Nonmalignant Lung Disease. *Cancer Research*. 2009;69(24):9202-9210. doi:10.1158/0008-5472.CAN-09-1378.

114. Silvestri GA, Vachani A, Whitney D, et al. A Bronchial Genomic Classifier for the Diagnostic Evaluation of Lung Cancer. *New England Journal of Medicine.* 2015;373(3):243-251. doi:10.1056/NEJMoa1504601.

115. Silvestri GA, Vachani A, Whitney D, et al. A Bronchial Genomic Classifier for the Diagnostic Evaluation of Lung Cancer. *New England Journal of Medicine.* 2015;373(3):243-251. doi:10.1056/NEJMoa1504601.

116. Song L, Langfelder P, Horvath S. Comparison of co-expression measures: mutual information, correlation, and model based indices. *BMC Bioinformatics*. 2012;13(1):328. doi:10.1186/1471-2105-13-328.

117. Sorensen BS, Wu L, Wei W, et al. Monitoring of epidermal growth factor

receptor tyrosine kinase inhibitor-sensitizing and resistance mutations in the plasma DNA of patients with advanced non–small cell lung cancer during treatment with erlotinib. *Cancer*. 2014;120(24):3896-3901. doi:10.1002/cncr.28964.

118. Sozzi G, Boeri M, Rossi M, et al. Clinical utility of a plasma-based miRNA signature classifier within computed tomography lung cancer screening: a correlative MILD trial study. *Journal of Clinical Oncology*. 2014;32(8):768-773. doi:10.1200/JCO.2013.50.4357.

119. Spira A, Beane J, Shah V, et al. Effects of cigarette smoke on the human airway epithelial cell transcriptome. *Proceedings of the National Academy of Sciences of the United States of America*. 2004;101(27):10143-10148. doi:10.1073/pnas.0401422101.

120. Spira A, Beane JE, Shah V, et al. Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer. *Nature Medicine*. 2007;13(3):361-366. doi:10.1038/nm1556.

121. Sporn MB, Dunlop NM, Newton DL, Smith JM. Prevention of chemical carcinogenesis by vitamin A and its synthetic analogs (retinoids). *Federation Proceedings.* 1976;35(6):1332-1338.

122. Steiling K, van den Berge M, Hijazi K, et al. A dynamic bronchial airway gene expression signature of chronic obstructive pulmonary disease and lung function impairment. *Am J Respir Crit Care Med*. 2013;187(9):933-942.

123. Steiling K, van den Berge M, Sebastiani P, et al. Airway Gene Expression as a Molecular Phenotype of COPD. December 2012. doi:10.1513/pats.8.2.208.

124. Strimbu K, Tavel JA. What are biomarkers? *Current Opinion in HIV and AIDS*. 2010;5(6):463-466. doi:10.1097/COH.0b013e32833ed177.

125. Su Z, Fang H, Hong H, et al. An investigation of biomarkers derived from legacy microarray data for their utility in the RNA-seq era. *Genome Biology*. 2014;15(12):523. doi:10.1186/s13059-014-0523-y.

126. Subramanian A, Subramanian A, Tamayo P, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*. 2005;102(43):15545-15550. doi:10.1073/pnas.0506580102.

127. Szabo E. Altered histology provides a positive clinical signal in the bronchial epithelium. 2011;4(6):775-778. doi:10.1158/1940-6207.CAPR-11-0214.

128. Szpechcinski A, Rudzinski P, Kupis W, Langfort R, Orlowski T, Chorostowska-Wynimko J. Plasma cell-free DNA levels and integrity in patients with chest radiological findings: NSCLC versus benign lung nodules. *Cancer Letters*. 2016;374(2):202-207. doi:10.1016/j.canlet.2016.02.002.

129. Takano T, Fukui T, Ohe Y, et al. EGFRMutations Predict Survival Benefit From Gefitinib in Patients With Advanced Lung Adenocarcinoma: A Historical Comparison of Patients Treated Before and After Gefitinib Approval in Japan. *Journal of Clinical Oncology*. 2008;26(34):5589-5595. doi:10.1200/JCO.2008.16.7254.

130. Tammemagi MC, Tammemagi MC, Lam SC, et al. Incremental Value of Pulmonary Function and Sputum DNA Image Cytometry in Lung Cancer Risk Prediction. *Cancer Prevention Research*. 2011;4(4):552-561. doi:10.1158/1940-6207.CAPR-10-0183.

131. Trang P, Medina PP, Wiggins JF, et al. Regression of murine lung tumors by the let-7 microRNA. *Oncogene*. 2009;29(11):1580-1587. doi:10.1038/onc.2009.445.

132. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*. 2009;25(9):1105-1111. doi:10.1093/bioinformatics/btp120.

133. Trapnell C, Roberts A, Goff L, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols*. 2012;7(3):562-578. doi:10.1038/nprot.2012.016.

134. Trapnell C, Williams BA, Pertea G, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*. 2010;28(5):511-515. doi:10.1038/nbt.1621.

135. Udyavar AR, Hoeksema MD, Clark JE, et al. Co-expression network analysis identifies Spleen Tyrosine Kinase (SYK) as a candidate oncogenic driver in a subset of small-cell lung cancer. *BMC Systems Biology*. 2013;7 Suppl 5:S1-S1. doi:10.1186/1752-0509-7-S5-S1.

136. van Boerdonk RAA, Smesseim I, Heideman DAM, et al. Close Surveillance

with Long-Term Follow-up of Subjects with Preinvasive Endobronchial Lesions. *American Journal of Respiratory and Critical Care Medicine*. 2015;192(12):1483-1489. doi:10.1164/rccm.201504-0822OC.

137. Voet D, Jing R, Cibulskis K, et al. Comprehensive genomic characterization of squamous cell lung cancers. *Nature*. 2012;489(7417):519-525. doi:10.1038/nature11404.

138. Wang G, Wang R, Strulovici-Barel Y, et al. Persistence of Smoking-Induced Dysregulation of MiRNA Expression in the Small Airway Epithelium Despite Smoking Cessation. Yildirim AÖ, ed. *PLoS One*. 2015;10(4):e0120824. doi:10.1371/journal.pone.0120824.

139. Wang L, Nie J, Sicotte H, et al. Measure transcript integrity using RNA-seq data. *BMC Bioinformatics*. 2016;17(1):e1261. doi:10.1186/s12859-016-0922-z.

140. Wang L, Wang S, Li W. RSeQC: quality control of RNA-seq experiments. *Bioinformatics (Oxford, England)*. 2012;28(16):2184-2185. doi:10.1093/bioinformatics/bts356.

141. Whitney DH, Elashoff MR, Porta Smith K, et al. Derivation of a bronchial genomic classifier for lung cancer in a prospective study of patients undergoing diagnostic bronchoscopy. *BMC Medical Genomics*. 2015; 8(1):18. doi:10.1186/s12920-015-0091-3.

142. Xing L, Su J, Guarnera MA, et al. Sputum microRNA Biomarkers for Identifying Lung Cancer in Indeterminate Solitary Pulmonary Nodules. *Clin Cancer Res*. 2015;21(2):484-489. doi:10.1158/1078-0432.CCR-14-1873.

143. Xing L, Todd NW, Yu L, Fang H, Jiang F. Early detection of squamous cell lung cancer in sputum by a panel of microRNA markers. *Modern Pathology*. 2010;23(8):1157-1164. doi:10.1038/modpathol.2010.111.

144. Yip AM, Horvath S. The Generalized Topological Overlap Matrix for Detecting Modules in Gene Networks. *BIOCOMP*. 2006.

145. YOU Z, ZHOU Y, GUO Y, CHEN W, CHEN S, WANG X. Activating transcription factor 2 expression mediates cell proliferation and is associated with poor prognosis in human non-small cell lung carcinoma. *Oncology Letters*. 2016;11(1):760-766. doi:10.3892/ol.2015.3922.

146. Zander T, Hofmann A, Staratschek-Jox A, et al. Blood-Based Gene

Expression Signatures in Non-Small Cell Lung Cancer. *Clinical Cancer Research.* 2011;17(10):3360-3367. doi:10.1158/1078-0432.CCR-10-0533.

147. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*. 2005;4(1):Article17. doi:10.2202/1544-6115.1128.

148. Zhao J, Fu W, Liao H, et al. The regulatory and predictive functions of miR-17 and miR-92 families on cisplatin resistance of non-small cell lung cancer. *BMC Cancer*. 2015;15(1):114. doi:10.1186/s12885-015-1713-z.

149. Zhao L, Zhu J, Zhou H, et al. Identification of cellular microRNA-136 as a dual regulator of RIG-I-mediated innate immunity that antagonizes H5N1 IAV replication in A549 cells. *Scientific Reports*. 2015;5:14991. doi:10.1038/srep14991.

150. Zhou T, Zhang W, Sweiss NJ, et al. Peripheral Blood Gene Expression as a Novel Genomic Biomarker in Complicated Sarcoidosis. Morty RE, ed. *PLoS One*. 2012;7(9):e44818. doi:10.1371/journal.pone.0044818.

151. Zhu K, Ding H, Wang W, et al. Tumor-suppressive miR-218-5p inhibits cancer cell proliferation and migration via EGFR in non-small cell lung cancer. *Oncotarget*. 2016;7(19):28075-28085. doi:10.18632/oncotarget.8576.

152. Zhu W-Y, Luo B, An J-Y, et al. Differential expression of miR-125a-5p and let-7e predicts the progression and prognosis of non-small cell lung cancer. *Cancer Invest*. 2014;32(8):394-401. doi:10.3109/07357907.2014.922569.

CURRICULUM VITAE